Instituto Nacional de Estadística

Oposiciones al Cuerpo de Diplomados en Estadística del Estado

BOE núm. 270, de 12 de octubre de 2020, pág. 87165

Estadística

INE

28 de septiembre de 2021

Índice general

Tema 16

Estadística descriptiva V. Momentos. Cálculo y aplicaciones. Medidas de simetría y curtosis. Medidas de concentración. Índice de Gini. Curva de Lorenz.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

Venancio Tomeo Perucha e Isaías Uña Juárez (2009). *Estadística descriptiva*. Madrid: Ibergarceta Publicaciones

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

16.1 Introducción

Una primera aproximación al estudio de una distribución, además del cálculo de las medidas de posición y dispersión, es su representación gráfica. En este tema analizaremos las distintas formas que pueden presentar las distribuciones comparándolas con un modelo ideal. El modelo de referencia será la distribución normal.

En cuanto a la forma de una distribución nos podemos hacer diversas preguntas. Por ejemplo, ¿es la distribución simétrica con respecto a un eje vertical $x = 1$?, ¿la mayoría de los valores se encuentran situados en torno al valor 3?...

Supongamos que con motivo del inicio del curso escolar, se desea conocer la distribución del uso del ordenador de los estudiantes de formación profesional. Para ello, se hace una representación de dicha distribución que es la que se presenta a continuación.

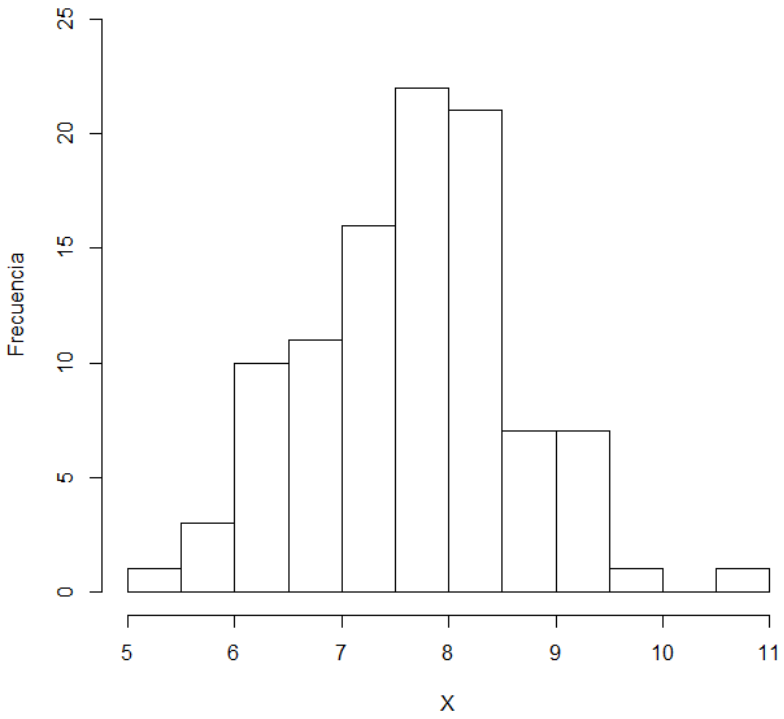


Figura 16.1: Histograma del uso del ordenador (horas)

A la vista de la gráfica, nos podemos preguntar si la distribución simétrica con respecto a un eje vertical situado en $x = 7$?, o si la mayoría de los valores se encuentran situados en torno al valor ese valor.

En definitiva, en este tema nos haremos preguntas a cerca de la simetría y la curtosis de una distribución. Además de estas cuestiones, estudiaremos la concentración de una distribución estadística.

16.2 Momentos

Dada una distribución estadística unidimensional, existen una serie de valores que la caracterizan y que denominamos **momentos**. Los momentos más conocidos son la media y la varianza. También nos podemos encontrar otros momentos en distintas medidas de centralización, dispersión, forma, etc.

Los momentos se clasifican en dos grandes grupos: momentos ordinarios y momentos centrales.

16.2.1 Momentos ordinarios o respecto al origen

Se llama momento ordinario (o respecto al origen) de orden p, a la media aritmética de las potencias de orden p de los valores de la variable. Dada una variable estadística X, el momento ordinario de orden p se denota $a_p(X)$ y se define como

$$a_p(X) = \frac{\sum_{i=1}^{I} x_i^p n_i}{N}.$$

Algunos casos particulares de los momentos ordinarios son:

- Momento ordinario de orden 0: $p = 0$ y $a_0(X) = \frac{\sum_{i=1}^{I} x_i^0 n_i}{N} = 1$.
- Momento ordinario de orden 1: $p = 1$ y $a_1(X) = \frac{\sum_{i=1}^{I} x_i^1 n_i}{N} = \bar{x}$.
- Momento ordinario de orden 2: $p = 0$ y $a_2(X) = \frac{\sum_{i=1}^{I} x_i^2 n_i}{N}$.

La combinación de los momentos de orden uno y dos se emplean en el cálculo de la varianza: $S^2 = a_2(X) - a_1^2(X)$.

Ejemplo 1. El departamento de calidad de una empresa de muebles ha detectado distintos defectos a lo largo de 100 tablas:

x_i (Número de defectos)	n_i
1	40
2	15
3	7
4	10
5	5
6	9
7	3
8	1
9	10

Para el cálculo de los momentos ordinarios uno y dos, añadiremos dos columnas a la tabla anterior.

x_i (Número de defectos)	n_i	$x_i n_i$	$x_i^2 n_i$
1	40	40	40
2	15	30	60
3	7	21	63
4	10	40	160
5	5	25	125
6	9	54	324
7	3	21	147
8	1	8	64
9	10	90	810
	$N = 100$	$\sum = 329$	$\sum = 1793$

El momento ordinario de orden 0 es:

$$a_0(X) = \frac{\sum_{i=1}^{I} x_i^0 n_i}{N} = \frac{1^0 * 40 + 2^0 * 15 + ... + 9^0 * 10}{100} = \frac{100}{100} = 1.$$

El momento ordinario de orden 1 es:

$$a_1(X) = \frac{\sum_{i=1}^{I} x_i^1 n_i}{N} = \frac{1 * 40 + 2 * 15 + ... + 9 * 10}{100} = \frac{329}{100} = 3,29.$$

Por tanto,

$$\bar{x} = 3,29.$$

El momento ordinario de orden 2 es:

$$a_2(X) = \frac{\sum_{i=1}^{I} x_i^2 n_i}{N} = \frac{1^2 * 40 + 2^2 * 15 + ... + 9^2 * 10}{100} = \frac{1793}{100} = 17,93.$$

16.2.2 Momentos centrales o respecto a la media

Se llama momento central (o respecto a la media) de orden p, a la media aritmética de las potencias de orden p de los valores de la variable menos la media aritmética (desviaciones respecto a la media). Dada una variable estadística X, el momento central de orden p se denota $m_p(X)$ y se define como

$$m_p(X) = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^p n_i}{N}$$

Algunos casos particulares de los momentos centrales son:

- Momento central de orden 0: $p = 0$ y $m_0(X) = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^0 n_i}{N} = 1$.

- Momento central de orden 1: $p = 1$ y $m_1(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^1 n_i}{N} = 0$.
- Momento central de orden 2: $p = 0$ y $m_2(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^2 n_i}{N} = S^2$.

Ejemplo 2. Retomando los datos del Ejemplo 1, vamos a calcular los momentos centrales hasta el orden 2:

El momento central de orden 0 es:

$$m_0(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^0 n_i}{N} = \\ = \frac{(1 - 3,29)^0 * 40 + (2 - 3,29)^0 * 15 + ... + (9 - 3,29)^0 * 10}{100} = 1.$$

El momento central de orden 1 es:

$$m_1(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^1 n_i}{N} = \\ = \frac{(1 - 3,29)^1 * 40 + (2 - 3,29)^1 * 15 + ... + (9 - 3,29)^1 * 10}{100} = 0.$$

El momento central de orden 2 es:

$$m_2(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^2 n_i}{N} = \\ = \frac{(1 - 3,29)^2 * 40 + (2 - 3,29)^2 * 15 + ... + (9 - 3,29)^2 * 10}{100} = \\ = \frac{710,59}{100} = 7,1059.$$

Por tanto,

$$S^2 = 7,1059.$$

Recordemos que varianza también la podemos calcular como

$$S^2 = a_2(X) - a_1^2(X) = 17,93 - 3,29^2 = 7,1059.$$

16.2.3 Momentos respecto a un punto

Se llama momento respecto al punto C de orden p, a la media aritmética de las potencias de orden p de los valores de la variable menos el valor C (desviaciones respecto al punto C). Dada una variable estadística X, el momento respecto al punto C de orden p se denota $v_p(X)$ y se define como

$$v_p(X) = \frac{\sum_{i=1}^{I}(x_i - C)^p n_i}{N}$$

Algunos casos particulares de los momentos respecto a un punto son:

- Momento respecto a un punto $C = 0$: $v_p(X) = \frac{\sum_{i=1}^{I}(x_i - 0)^p n_i}{N}$. Este caso coincide con la definición de los momentos ordinarios.
- Momento respecto a un punto $C = \bar{x}$: $m_p(X) = \frac{\sum_{i=1}^{I}(x_i - \bar{x})^p n_i}{N}$. Este caso coincide con la definición de los momentos centrales.

Ejemplo 3. Retomando los datos del Ejemplo 1, vamos a calcular los momentos respecto al punto 2 de orden 1 y 2:

$$\begin{aligned} v_1(X) &= \frac{\sum_{i=1}^{I}(x_i - 2)^1 n_i}{N} = \\ &= \frac{(1-2) * 40 + (2-2) * 15 + ... + (9-2) * 10}{100} = \\ &= \frac{129}{100} = 1,29. \end{aligned}$$

$$\begin{aligned} v_2(X) &= \frac{\sum_{i=1}^{I}(x_i - 2)^2 n_i}{N} = \\ &= \frac{(1-2)^2 * 40 + (2-2)^2 * 15 + ... + (9-2)^2 * 10}{100} = \\ &= \frac{877}{100} = 8,77. \end{aligned}$$

16.2.4 Relaciones entre momentos

Podemos establecer relaciones entre los momentos ordinarios y los momentos de orden central. Recurriendo al binomio de Newton y al número combinatorio, tenemos la siguiente relación:

$$m_p = \sum_{k=0}^{p}(-1)^k \cdot \binom{p}{k} \cdot a_1^k \cdot a_{p-k},$$

donde:

$$\binom{p}{k} = \frac{p!}{k!(p-k)!}.$$

Las relaciones más utilizadas son las de los momentos centrales hasta el orden cuatro. Por ejemplo, las relaciones entre los momentos de orden tres y orden dos son:

$$m_2(X) = a_2(X) - a_1^2(X)$$

$$m_3(X) = a_3(X) - 3a_1(X)a_2(X) + 2a_1^3(X)$$

Ejemplo 4. Retomando los datos del Ejemplo 1, calcula el momento central de orden tres a partir de los momentos ordinarios.

Para el cálculo de los momentos ordinarios uno, dos y tres, añadiremos tres columnas.

x_i (Número de defectos)	n_i	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$
1	40	40	40	40
2	15	30	60	120
3	7	21	63	189
4	10	40	160	640
5	5	25	125	625
6	9	54	324	1944
7	3	21	147	1029
8	1	8	64	512
9	10	90	810	7290
	$N = 100$	$\sum = 329$	$\sum = 1793$	$\sum = 12389$

$a_1(X) = \frac{329}{100} = 3,29$

$a_2(X) = \frac{1793}{100} = 17,93$

$a_3(X) = \frac{12389}{100} = 123,89$

Por tanto,

$$m_3(X) = a_3(X) - 3a_1(X)a_2(X) + 2a_1^3(X) = 18,1435.$$

16.3 Medidas de forma

La forma de una distribución se suele comparar con la distribución Normal. Ésta distribución tiene forma de campana. Las comparaciones se hacen sobre dos aspectos fundamentales: simetría y curtosis.

16.3.1 Medidas de simetría

La simetría de una distribución puede obtenerse respecto de un valor. En nuestro caso, estudiaremos la simetría respecto a la media. Es decir, diremos que una distribución es simétrica respecto a la media si al "doblar" la distribución por el eje de simetría ambas partes coinciden.

En la siguiente figura se ha tomando como ejemplo la distribución normal estándar. Esta distribución es simétrica respecto del valor de su media. Y a su vez la media coincide con la mediana y la moda en el valor cero (es decir, $\bar{x} = M_e = M_o = 0$).

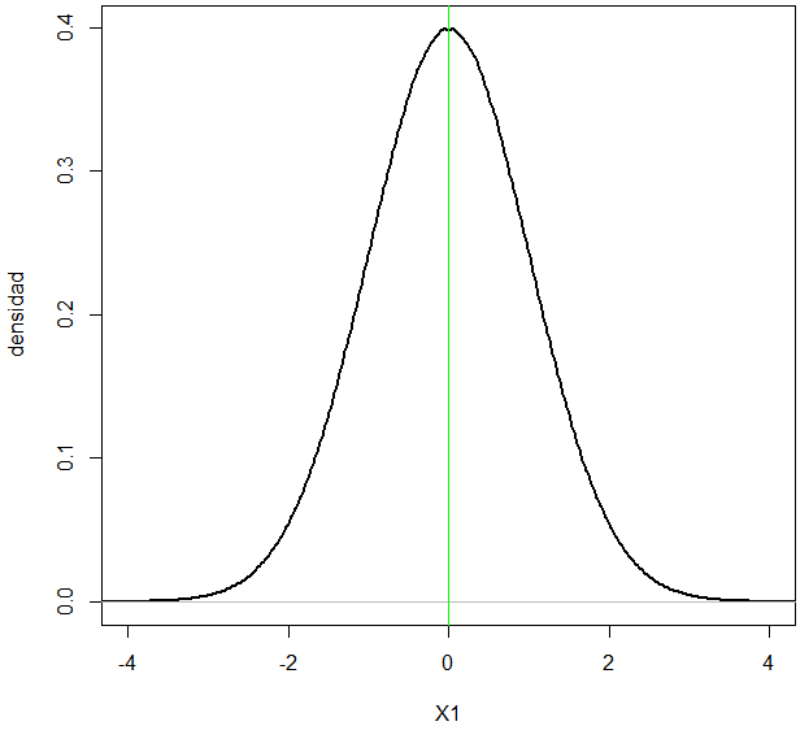


Figura 16.2: Distribución normal estándar

Para medir la asimetría de una distribución se utiliza el coeficiente de asimetría de Fisher. Este coeficiente se calcula como el cociente entre el momento central de orden tres y la desviación típica elevada al cubo. Se denota por $g_1(X)$ y la expresión del coeficiente

de asimetría viene dado por:

$$g_1(X) = \frac{m_3(X)}{S^3} = \frac{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^3 * n_i}{N}}{S^3}.$$

Cuando los valores son unitarios la expresión del coeficiente de asimetría se puede simplificar del siguiente modo:

$$g_1(X) = \frac{m_3(X)}{S^3} = \frac{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^3}{N}}{S^3}.$$

Tipos de asimetría

- Distribución asimétrica negativa

  Si $g_1(X) < 0$, la distribución presenta asimetría negativa. Es decir, los valores altos son los más frecuentes y tiene una cola a la izquierda. En la siguiente figura se muestra un ejemplo de una distribución asimétrica negativa.

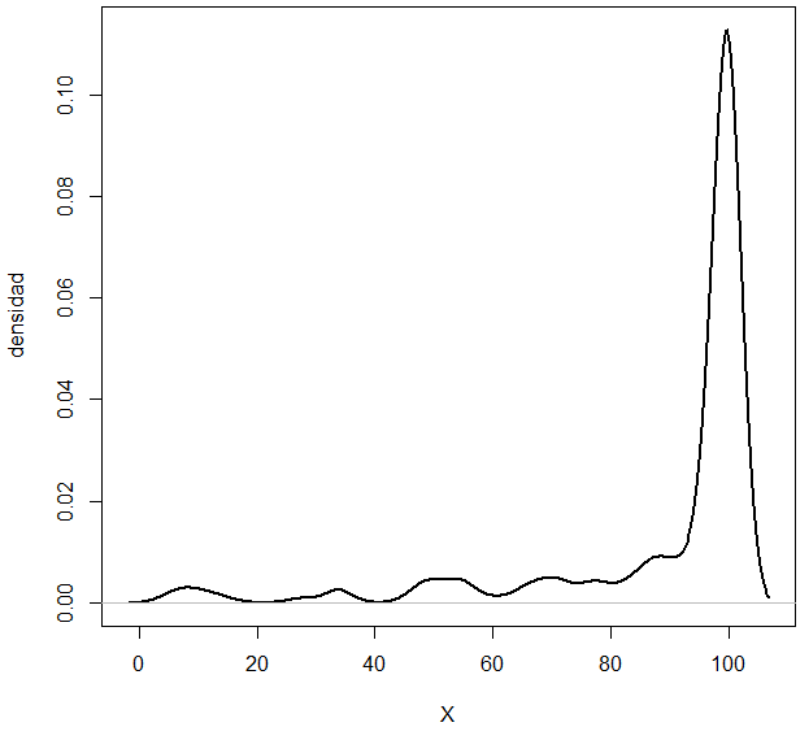


Figura 16.3: Distribución asimétrica negativa

- Distribución simétrica

  Si una distribución es simétrica, su coeficiente de simetría será cero. Es decir,

  $$X \text{ simétrica } \rightarrow g_1(X) = 0.$$

  En la siguiente figura se muestra un ejemplo de una distribución simétrica.

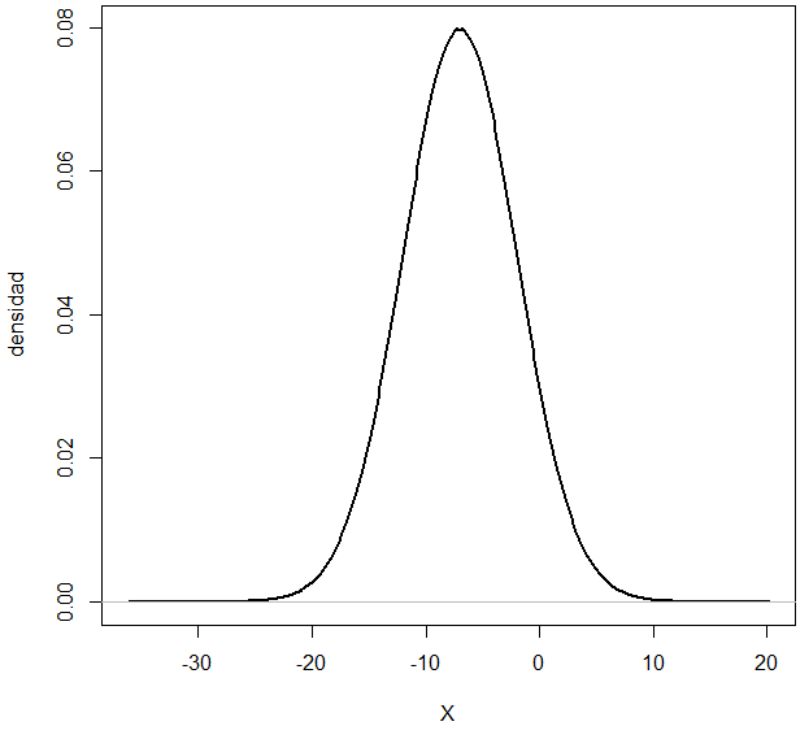


Figura 16.4: Distribución simétrica

  Se debe tener en cuenta que cuando el coeficiente de asimetría toma el valor de cero, esto no implica que la distribución sea simétrica. Puede ser o no simétrica. En la práctica, se representa la distribución para poder confirmar su simetría.

- Distribución asimétrica positiva

  Si $g_1(X) > 0$, la distribución presenta asimetría positiva. Es decir, los valores bajos son los más frecuentes y tiene una cola a la derecha. En la siguiente figura se muestra un ejemplo de una distribución asimétrica positiva.

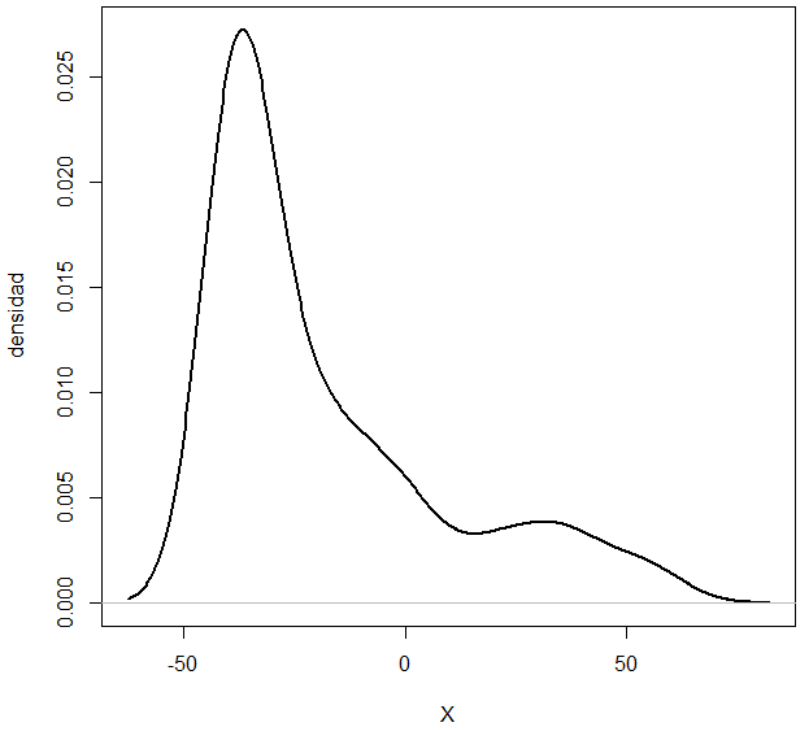


Figura 16.5: Distribución asimétrica positiva

Ejemplo 5. Se han puntuado 105 empresas en función a su retraso en los pagos a sus proveedores. Se desea saber si la distribución es simétrica para informar a los inversores extranjeros.

L_{i-1}, L_i	n_i
[0, 3)	7
[3, 4)	10
[4, 6)	35
[6, 7)	50
[7, 10]	3

Para calcular el coeficiente de asimetría hay que obtener la marca de clase.

L_{i-1}, L_i	n_i	x_i
[0, 3)	7	1,5
[3, 4)	10	3,5
[4, 6)	35	5
[6, 7)	50	6,5
[7, 10]	3	8,5

Y utilizando la forma directa,

$$g_1(X) = \frac{m_3(X)}{S^3} = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 * n_i}{N}}{S^3} = \frac{-3,4026}{1,5035^3} = -1,0012.$$

Dado que el coeficiente de asimetría es inferior a cero (negativo), la distribución es simétrica negativa.

Alternativamente, podemos emplear la relación entre los momentos centrales y los momentos ordinarios.

L_{i-1}, L_i	x_i	n_i	$x_i * n_i$	$x_i^2 * n_i$	$x_i^3 * n_i$
[0, 3)	1.5	7	10,500	15,750	23,625
[3, 4)	3.5	10	35,000	122,500	428,750
[4, 6)	5	35	175,000	875,000	4375,000
[6, 7)	6.5	50	325,000	2112,500	13731,250
[7, 10]	8.5	3	25,500	216,750	1842,375
			$\sum = 571,000$	$\sum = 3342,500$	$\sum = 20401,000$

$a_1(X) = 5,4381$

$a_2(X) = 31,8333$

$a_3(X) = 194,2952$

$m_2(X) = a_2(X) - a_1^2(X) \rightarrow S = \sqrt{m_2(X)} = 1,5035$

$m_3(X) = a_3(X) - 3a_1(X)a_2(X) + 2a_1^3(X) = -3,4026$

Por tanto, $g_1(X) = \frac{m_3(X)}{S^3} = \frac{-3,4026}{1,5035^3} = -1,0012.$

Gráficamente, podemos observar que la distribución presenta una cola más larga a la izquierda del valor central.

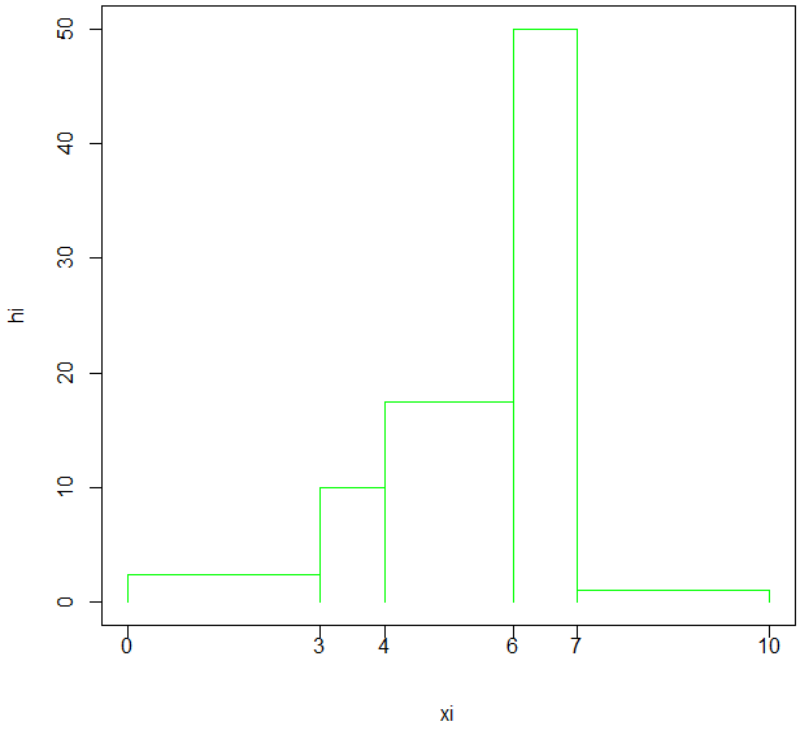


Figura 16.6: Histograma del retraso en pagos

Otras medidas de asimetría

Además del coeficiente de asimetría de Fisher, es posible utilizar otras medidas. Se pueden destacar:

- Coeficiente de asimetría de Pearson
- Coeficiente de sesgo cuartílico
- Coeficiente de sesgo percentílico

La interpretación de estos coeficientes es similar al caso del coeficiente de asimetría de Fisher. En la práctica, no es frecuente el empleo de los coeficientes de sesgo cuartílico y percentílico para el estudio de la asimetría.

En el caso del coeficiente de asimetría de Pearson, sólo se consideran las distribuciones unimodales y campaniformes para comparar la distancia de la media a la moda.

Si la media coincide con la moda, la distribución es simétrica. En la siguiente figura se cumple esta condición.

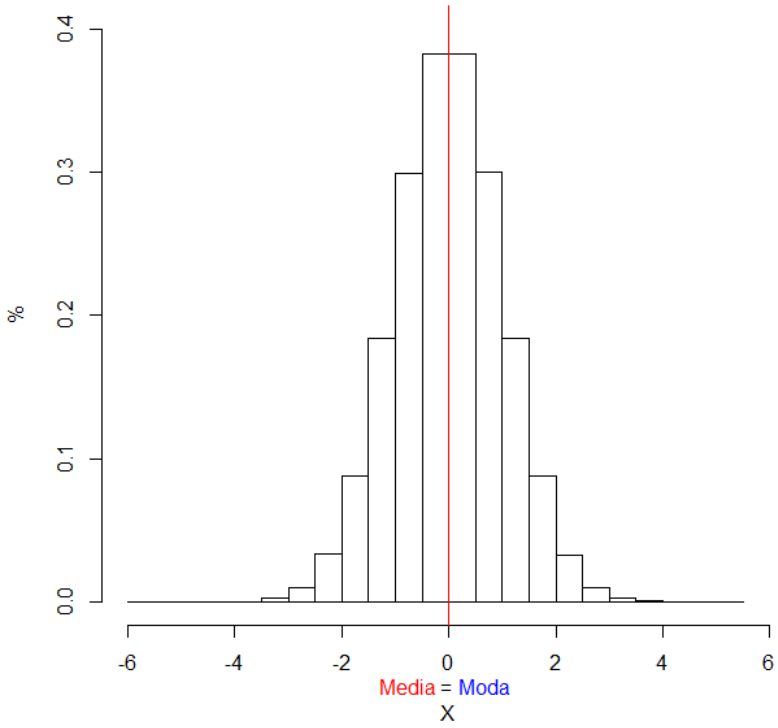


Figura 16.7: Distribución simétrica

Si la media es superior a la moda, la distribución es asimétrica positiva. En la siguiente figura se cumple esta condición.

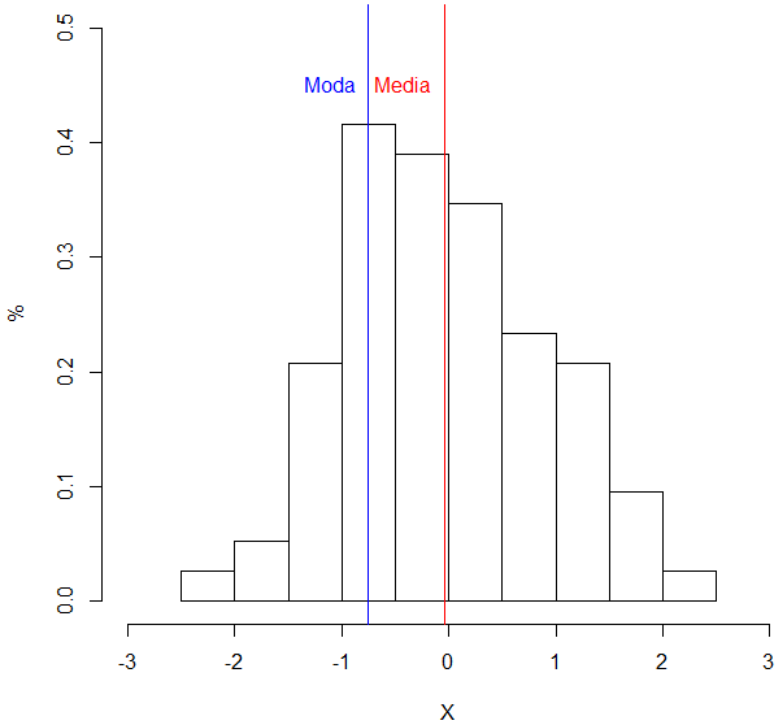


Figura 16.8: Distribución asimétrica positiva

Por último, si la media es inferior, es asimétrica negativa. En la siguiente figura se cumple esta condición.

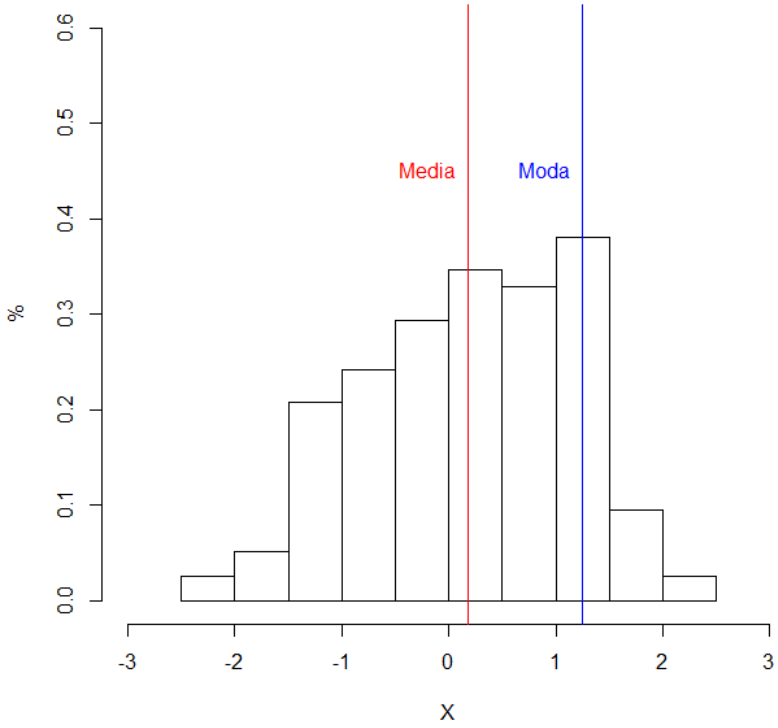


Figura 16.9: Distribución asimétrica negativa

16.3.2 Medidas de curtosis

Las medidas de apuntamiento(o curtosis) analizan la concentración de valores de la variable estadística en torno a la media artimética (que es nuestro punto de referencia para la simetría). Se toma como referencia la distribución Normal, y en base a ella se compara nuestra distribución que puede ser más apuntada, más achatada o similar a la Normal.

El grado de apuntamiento se mide a través del coeficiente de curtosis que se denota por g_2 y viene expresado como

$$g_2(X) = \frac{m_4(X)}{S^4} - 3 = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 * n_i}{N}}{S^4} - 3.$$

Cuando los valores son unitarios la expresión del coeficiente de curtosis se puede simplificar del siguiente modo:

$$g_2(X) = \frac{m_4(X)}{S^4} - 3 = \frac{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N}}{S^4} - 3.$$

El cociente entre momento central de orden cuatro y la desviación típica a la cuarta es igual a 3 si la distribución es Normal. En este sentido, se resta el valor 3 en el coeficiente de apuntamiento para comparar cualquier distribución con la distribución Normal. Esta versión del coeficiente de apuntamiento también se conoce como el *coeficiente de apuntamiento de Fisher*.

Tipos de distribuciones según su grado de apuntamiento

- Distribución leptocúrtica

  Si $g_2(X) > 0$, la distribución es leptocúrtica. Alto número de valores se concentran en torno a la media aritmética. Esta distribución es más apuntada que la distribución Normal.

- Distribución mesocúrtica

  Si $g_2(X) = 0$, la distribución es mesocúrtica. Esta distribución es igual de apuntada que la distribución Normal.

- Distribución platicúrtica

  Si $g_2(X) < 0$, la distribución es platicúrtica. Bajo número de valores se concentran en torno a la media aritmética. Esta distribución es más achatada que la distribución Normal.

  En la siguiente figura se muestra un ejemplo de cada tipo de distribución: leptocúrtica, $\chi^2 - 0,5$; mesocúrtica, $Z \sim$ N(0, 1); y platicúrtica, U(-3 , 3).

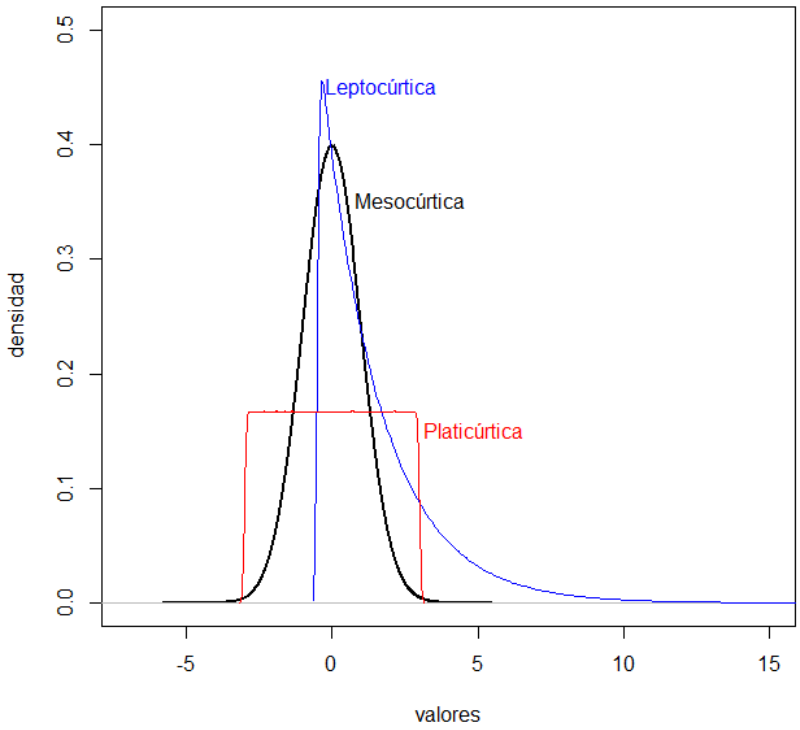


Figura 16.10: Tipos de apuntamiento

Ejemplo 6. Retomando los datos del Ejemplo 1, vamos a calcular el coeficiente de apuntamiento de la distribución estadística.

Utilizando la forma directa,

$$g_2(X) = \frac{m_4(X)}{S^4} - 3 = \frac{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 * n_i}{N}}{S^4} - 3 = \frac{133,6309}{2,6657^4} - 3 = -0,3535.$$

Dado que el coeficiente de curtosis es superior a cero (positivo), la distribución es leptocúrtica.

Alternativamente, podemos emplear la relación entre los momentos centrales y los momentos ordinarios.

x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$x_i^3 n_i$	$x_i^4 n_i$
1	40	40	40	40	40
2	15	30	60	120	240
3	7	21	63	189	567
4	10	40	160	640	2560
5	5	25	125	625	3125
6	9	54	324	1944	11664
7	3	21	147	1029	7203
8	1	8	64	512	4096
9	10	90	810	7290	65610
	$N = 100$	$\sum = 329$	$\sum = 1793$	$\sum = 12389$	$\sum = 95105$

$a_1(X) = 3,29$

$a_2(X) = 17,93$

$a_3(X) = 123,89$

$a_4(X) = 951,05$

$m_2(X) = a_2(X) - a_1^2(X) \rightarrow S = \sqrt{m_2(X)} = 2,6657$

$m_4(X) = a_4(X) - 4a_1(X)a_3(X) + 6a_1^2(X)a_2(X) - 3a_1^4(X) = 133,6309$

Por tanto,

$$g_2(X) = \frac{m_4(X)}{S^4} - 3 = \frac{133,6309}{2,6657^4} - 3 = -0,3535.$$

Gráficamente, podemos observar que la distribución presenta pocos valores en torno a la zona central.

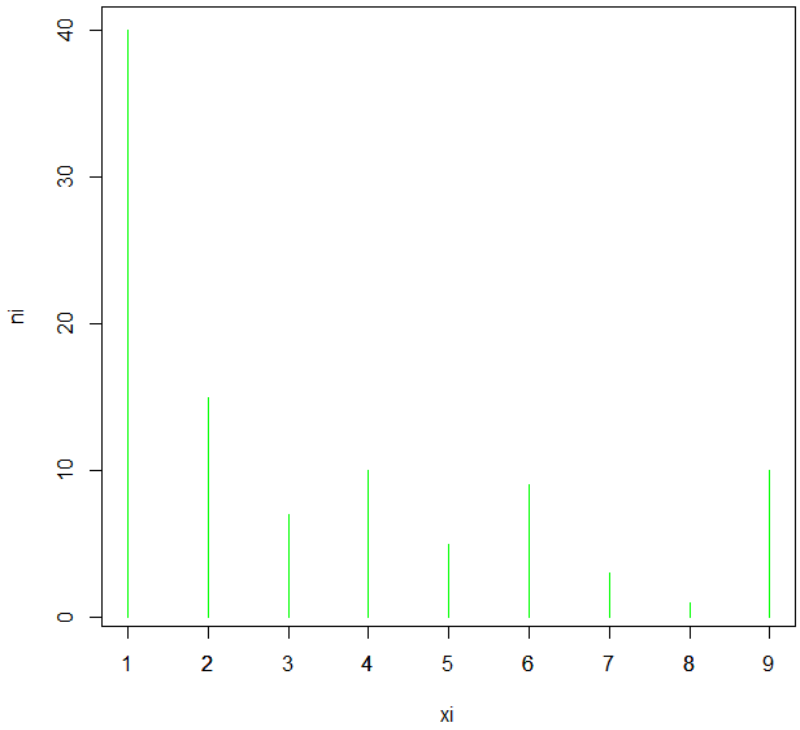


Figura 16.11: Diagrama de barras del número de defectos

Otras medidas de apuntamiento

Además del coeficiente de apuntamiento de Fisher, es posible utilizar otras medidas. Entre ellas destaca el coeficiente de apuntamiento percentílico. La interpretación de este coeficiente es similar al caso del coeficiente de apuntamiento de Fisher, con la diferencia de que se toma el valor 3 como punto de comparación (en lugar del cero).

16.4 Medidas de concentración

La concentración es un término económico que mide el grado de igualdad en el reparto de los valores de una variable estadística (o económica). No debe confundirse la concentración como un término opuesto a la dispersión. En el ámbito económico, puede entenderse como lo contrario al reparto equitativo.

Para medir la concentración se empleará el índice de concentración de Gini y, gráficamente, la curva de Lorenz.

16.4.1 Índice de Gini

El índice de Gini es una medida cuantitativa que mide la concentración y se denota por IG.

A nivel práctico se tienen en cuenta los valores de la variable (x_i), las frecuencias absolutas (n_i), y se calcula el volumen (o riqueza del grupo) como el producto de los valores de las variables por las frecuencias absolutas ($x_i n_i$). En la expresión del índice de Gini se emplean las frecuencias relativas acumuladas expresadas en tanto por ciento.

$$IG = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i},$$

donde:
$p_i = \frac{N_i}{N} * 100$,
$q_i = \frac{\sum_{j=1}^{i} x_j n_j}{\sum_{j=1}^{k} x_j n_j} * 100$.

En cuanto a la interpretación del índice de Gini:

- Toma valores entre 0 y 1. Es decir, $0 \leq IG \leq 1$. El término medio es 0,5.
- Si $IG = 0$, la concentración es mínima (o, en términos económicos, el reparto es equitativo). Cuanto más cercano a cero, menor grado de concentración.
- Si $IG = 1$, la concentración es máxima (o, en términos económicos, el reparto no es equitativo). Cuanto más cercano a 1, mayor grado de concentración.

16.4.2 Curva de Lorenz

La curva de Lorenz relaciona el frecuencias relativas acumuladas (p_i) con las frecuencias relativas acumuladas del volumen de la variable (q_i). La curva empieza en el punto (0, 0) y finaliza en el punto (100, 100), y se representa por debajo de la bisectriz del primer cuadrante. Siempre se cumple que $p_i \geq q_i \forall i$.

Representación de la curva de Lorenz:

- Si $IG = 0$, la curva de Lorenz presenta la siguiente gráfica.

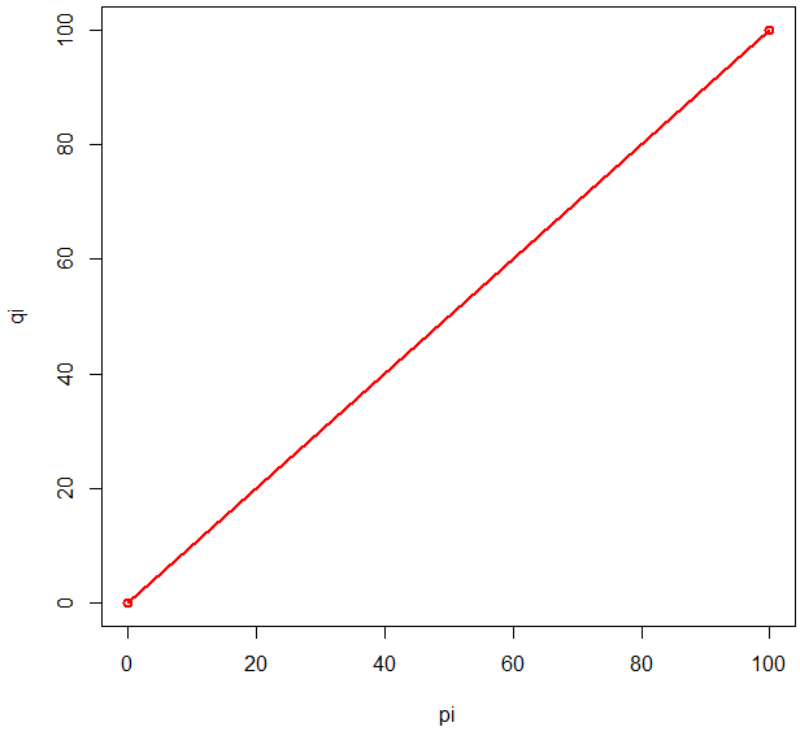


Figura 16.12: $IG = 0$

- Si $0 < IG < 1$, la curva de Lorenz presenta la siguiente gráfica.

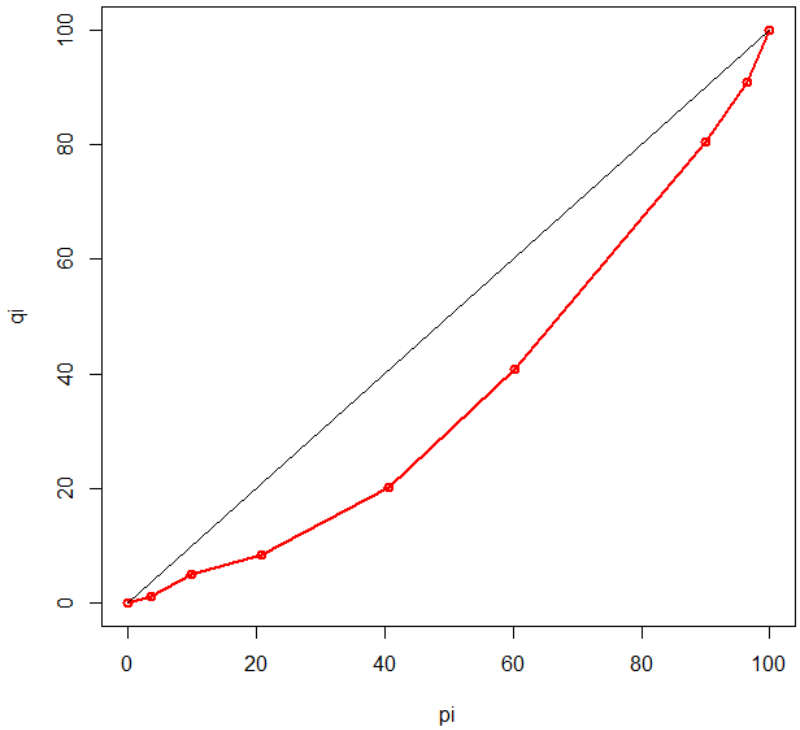


Figura 16.13: $0 < IG < 1$

- Si $IG = 1$, la curva de Lorenz presenta la siguiente gráfica.

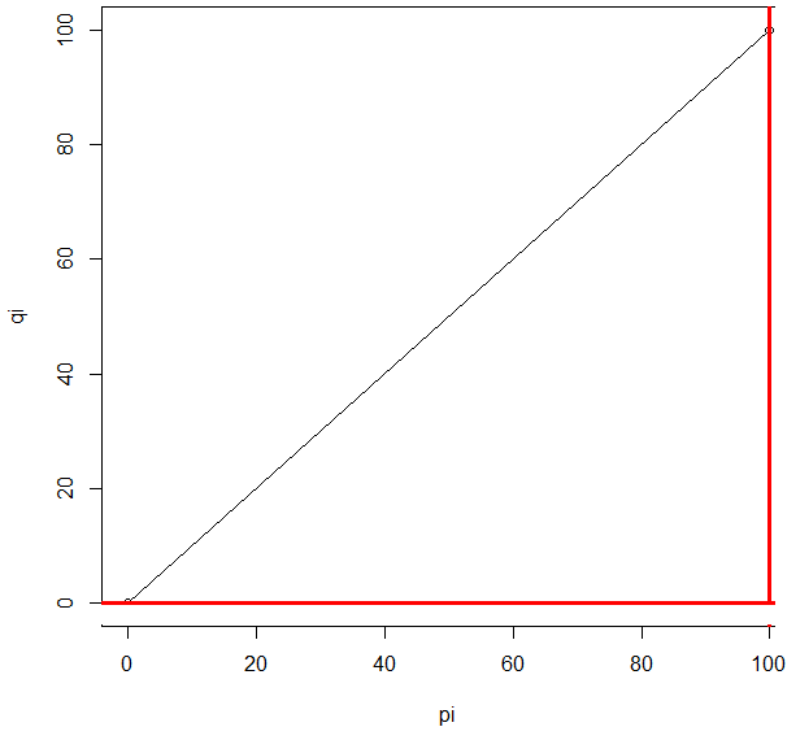


Figura 16.14: $IG = 1$

Ejemplo 7. Los estudiantes de último curso han informado a la comisión académica sobre sus ingresos brutos durante el curso 2016. La comisión desea conocer si el reparto de las ayudas económicas ha sido equitativo.

Ingresos en miles de euros (x_i)	Estudiantes (n_i)
3	4
4	6
5	10
7	20
11	22
12	31
16	15
24	4

Para medir la concentración, calcularemos el índice de Gini. Primero construiremos nuevas columnas en la tabla anterior.

x_i	n_i	$m_i = x_i n_i$	N_i	M_i	p_i	q_i	$p_i - q_i$
3	4	12	4	12	3,571	1,020	2,551
4	6	24	10	36	8,929	3,061	5,867
5	10	50	20	86	17,857	7,313	10,544
7	20	140	40	226	35,714	19,218	16,497
11	22	242	62	468	55,357	39,796	15,561
12	31	372	93	840	83,036	71,429	11,607
16	15	240	108	1080	96,429	91,837	4,592
24	4	96	112	1176	100,000	100,000	0,000

Por tanto, $IG = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{67{,}2194}{300{,}8929} = 0,2234$. La distribución de los ingresos (que parten de las ayudas académicas) de los distintos estudiantes presenta una baja concentración. Es decir, podemos considerar que las ayudas están relativamente bien repartidas en términos de equidad.

Gráficamente, podemos observar que la distribución presenta una concentración baja.

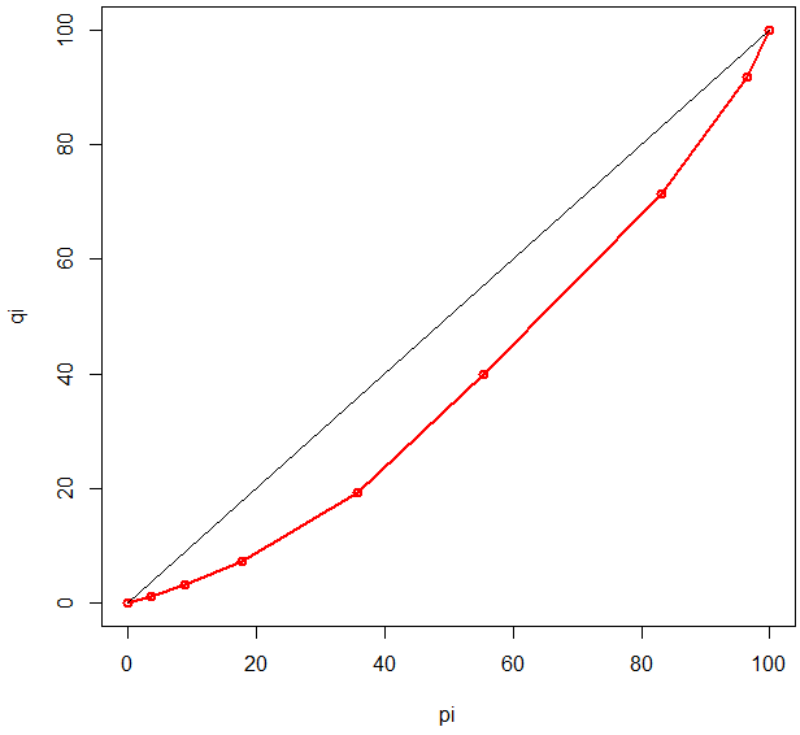


Figura 16.15: Concentración de los ingresos ($IG = 0,22$)

Problema propuesto. Un economista desea analizar la cotización de un grupo de empresas españolas. Para ello, se ha fijado en el precio de la acción en euros (X) y el efectivo en miles de euros (Y) en una sesión del Mercado Continuo. Los datos se presentan en la siguiente tabla:

X	n_i	Y	n_j
0,54	1	[2800, 5000)	3
30,7	6	[5000, 10000)	7
78,8	4	[10000, 15000)	6
153,5	8	[15000, 25000)	4
244,5	9	[25000, 40000)	7
559,4	5	[40000, 70000)	3
1300,2	2	[70000, 100000)	3
		[100000, 170000]	2

Se pide:

(a) Estudia la forma de la distribución de la cotización de las empresas.

(b) Estudia la concentración del efectivo. ¿Se puede considerar que el reparto es equitativo?

Por último, se recomienda la lectura de otros libros de varios autores para ampliar los conocimientos respecto de la materia estudiada (Bachero Nebot y col. 2006; Peña 2001).

Bibliografía

Peña, Daniel (2001). *Fundamentos de estadística*. Madrid: Alianza editorial (página 26).

Bachero Nebot, José, Olga Blasco Blasco, Vicente Coll Serrano, Rafael Diez García, Jesús Esteban García, Antonia Ivars Escortell, María Isabel Lopez Rodriguez, Concepción Rojo Olivas y Felix Ruiz Ponce (2006). *Estadística descriptiva y nociones de probabilidad*. Editorial Paraninfo (página 26).

Tomeo Perucha, Venancio e Isaías Uña Juárez (2009). *Estadística descriptiva*. Madrid: Ibergarceta Publicaciones (página 1).

Tema 17

Estadística descriptiva VI. Distribuciones estadísticas bidimensionales. Distribuciones marginales y condicionales. Independencia y asociación de las variables. Representación gráfica. Momentos en las distribuciones bidimensionales. Concepto de covarianza. Correlación. Significado.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

AM Montiel Torres, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson
D Peña (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial

17.1 Introducción

Hasta ahora hemos analizado cada variable de forma independiente, se analizaba por un lado la altura de los estudiantes del Grado en Estudios Estadísticos y por otro su peso. Sin embargo, en ocasiones es interesante analizar las dos características de la población de forma conjunta.

En este tema para cada individuo se observan dos características de forma conjunta, por ejemplo (Altura, Peso), (Gastos, Ingresos), (Tiempo de estudio, Notas), etc.

La **variable bidimensional** se va a denotar por (X, Y) y es el conjunto de datos (x_i, y_j) que indican las características de los individuos de la muestra o la población. Siendo X una variable unidimensional que toma los valores $x_1, x_2, ..., x_I$, con $I =$ al número de categorías o clases de la variable X e Y la variable unidimensional que toma los valores $y_1, y_2, ..., y_J$, siendo J el número de categorías de la variable Y.

Un ejemplo de variable estadística bidimensional es:

Individuo	Peso (kg.)	Altura (m.)
1	75	1,65
2	68	1,72
...	...	...
N	72	1,59

Tabla 17.1: Tabla de datos de la variable bidimensional $(X, Y) = (Peso, Altura)$

A las variables estadísticas X e Y se les denominan **variables marginales** y pueden ser ambas cuantitativas, ambas cualitativas o una de cada tipo; a su vez, los caracteres cuantitativos puede ser variables estadísticas tanto discretas como continuas.

Al igual que se hacía con las variables estadísticas unidimensionales, lo primero que se hace con los datos de las variables bidimensionales es agruparlos en **tablas estadísticas**. A partir de los datos de las tablas estadísticas se obtendrán distintas características de la variable (X, Y) que nos indicaran si las variables unidimensionales que la componen son independientes o están relacionadas.

17.2 Distribuciones estadísticas bidimensionales

Una forma de definir una variable estadística bidimensional es a partir de la tabla de frecuencias.

17.2.1 Distribución de frecuencias absolutas

Consideremos una población de N individuos sobre los que medimos conjuntamente dos variables unidimensionales, X e Y. Cada individuo vendrá dado entonces por un par de valores (x_i, y_i), $i = 1, ..., N$.

Individuo	x_i	y_i
1	x_1	y_1
2	x_2	y_2
...	...	...
N	x_N	y_N

Tabla 17.2: Tabla de datos de la variable bidimensional (X, Y)

A partir de la matriz de datos de la variable bidimensional (X, Y) se va a construir una tabla de doble entrada con tantas filas como categorías tenga la variable X y tantas columnas como categorías tenga la variable Y.

El valor de cada celda (i, j) se le denota por n_{ij} es la **frecuencia absoluta** de la categoría (x_i, y_j), es decir el número de individuos de toman el valor x_i para la variable X y simultáneamente toman el valor y_j para la variable Y.

Valores X \Y	y_1	y_2	...	y_j	...	y_J
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}
...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}
...	...	...	...	...	...	...
x_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}

Tabla 17.3: Tabla de frecuencias bidimensional

Si sumamos las frecuencias absolutas para todas las categorías de la variable (X, Y) obtendremos el número de individuos de la población o de la encuesta.

$N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$

17.2.2 Distribución de frecuencias relativas

Las frecuencia absoluta n_{ij} de la variable (X, Y) varía entre 0 y el tamaño de la muestra o la población N, es decir depende del tamaño de la población.
Para conocer si una categoría (x_i, y_j) tiene más o menos representatividad en la población se utiliza la **frecuencia relativa**.

La frecuencia relativa se calcula como la frecuencia absoluta dividida por el total de observaciones, N.

$$f_{ij} = \frac{n_{ij}}{N}$$

Ahora la frecuencia relativa varía entre 0 y 1, no depende del número de observaciones.

Valores X \Y	y_1	y_2	...	y_j	...	y_J
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1J}
x_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2J}
...	...	...	...	...	...	...
x_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{iJ}
...	...	...	...	...	...	...
x_I	f_{I1}	f_{I2}	...	f_{Ij}	...	f_{IJ}

Tabla 17.4: Tabla de frecuencias relativas bidimensional

Una propiedad de las frecuencias relativas es que la suma es 1.

$\sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}}{N} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \frac{N}{N} = 1$

Ejemplo 1. En la facultad de Estudios Estadísticos se ha preguntado a 95 estudiantes por el curso más alto en el que están matriculados y su edad, obteniendo los siguientes resultados:

Edad \Curso	Primero	Segundo	Tercero	Cuarto
18	20	0	0	0
19	6	15	0	0
20	5	8	12	0
≥ 21	4	5	8	12

Tabla 17.5: Tabla de frecuencias absolutas

se pide que construyas la tabla de frecuencias relativas para la variable conjunta.

Para construir la tabla de frecuencias relativas lo único que hay que hacer es dividir por el número de individuos la frecuencia absoluta de ca categoría.

Edad \Curso	Primero	Segundo	Tercero	Cuarto
18	0,21	0	0	0
19	0,06	0,16	0	0
20	0,05	0,085	0,13	0
≥ 21	0,04	0,05	0,085	0,13

Tabla 17.6: Tabla de frecuencias absolutas

Las tablas de frecuencias que se han construido es para variables que tienen un número finito de categorías. Si una de las variables es continua, el primer paso es agruparla en clases. Un ejemplo de tabla de frecuencias para variables continuas es la siguiente:

Peso \Altura	[1,5;1,60)	[1,6;1,70)	[1,7;1,80)	[1,8;1,90]
[55;70)	20	10	1	0
[70;85)	6	15	10	4
[85;100)	5	8	12	21

Tabla 17.7: Tabla de frecuencias absolutas

17.3 Representación gráfica

El principal gráfico utilizado cuando las dos variables, X e Y, son numéricas para comprobar si dos variables están relacionadas es el **diagrama de dispersión o nube de puntos**.

Si las variables son continuas, es mejor utilizar los datos de los individuos que las tablas de frecuencias donde ya se han agrupado en k categorías.

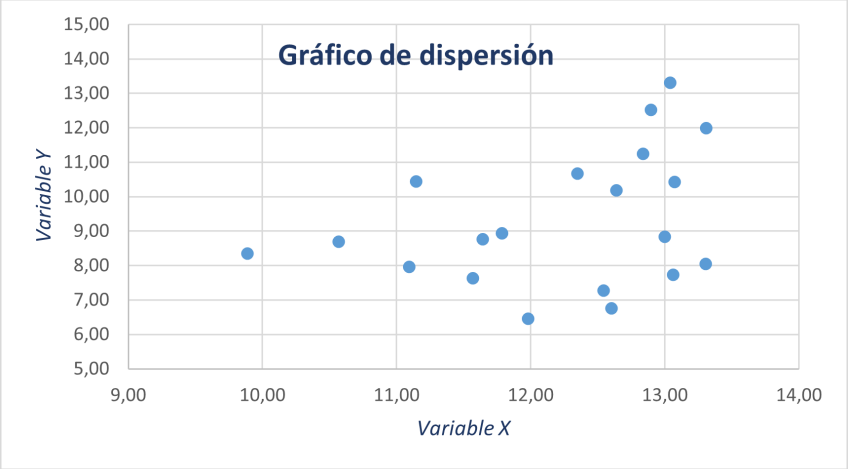


Figura 17.1: Digrama de dispersión o nube de puntos

El diagrama de dispersión no nos da la información del número de individuos que hay en cada punto, si el número de categorías de alguna de las variables es muy pequeño se recomienda el gráfico de barras.

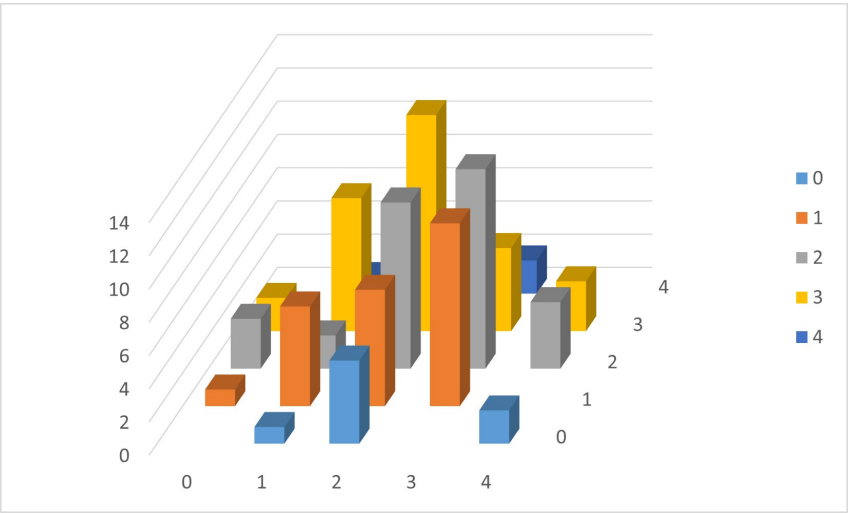


Figura 17.2: Diagrama de barras para la variable (X, Y)

En la Figura 17.2 se representa la variable conjunta (X, Y), siendo la altura de las barras la frecuencia n_{ij}. Cada color representa una variable condicionada y si existiese independencia todas las variables condicionadas deberían tener la misma forma.

17.4 Distribuciones marginales y condicionales

Aunque para cada individuo tenemos los valores tanto de la variable X como de la variable Y. También se puede realizar el estudio de cada una de las variables que componen la variable bidimensional, **distribuciones marginales**. Otro estudio interesante es el de una de las variables pero solamente para las observaciones que toman un determinado valor en la otra variable **variables condicionada**.

17.4.1 Distribuciones marginales

Las varibles marginales son cada una de las variables unidimensionales que componen la variable bidimensional.

Valores X \Y	y_1	y_2	...	y_j	...	y_J	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...	...
x_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
$n_{.j}$	$n_{,1}$	$n_{,2}$	...	$n_{.j}$	...	$n_{.J}$	N

Tabla 17.8: Frecuencias marginales

En la Tabla 17.8, la ultima columna corresponde a las frecuencias absolutas de la variable X y la última fila representa las frecuencias absolutas de la variable Y.

La variable marginal X toma los valores $x_1, x_2, ..., x_I$ (los mismos que tomaba cuando formaba parte de la variable bidimensional (X, Y)).

La frecuencia absoluta para la categoría x_i se denota por $n_{i.}$ e indica el número de individuos que toman el valor x_i para la variable X, independientemente del valor que toman para la variable Y.

$n_{i.} = \sum_{j=1}^{J} n_{ij}$ y $f_{i.} = \sum_{j=1}^{J} f_{ij} = \frac{n_{i.}}{N}$

X	$n_{i.}$	$f_{i.}$
x_1	$n_{1.}$	$f_{1.}$
x_2	$n_{2.}$	$f_{2.}$
...	...	...
x_i	$n_{i.}$	$f_{i.}$
...	...	...
x_I	$n_{I.}$	$f_{I.}$
Suma	N	1

Tabla 17.9: Frecuencias marginales de la variable X

Para la variable Y se hace lo mismo cambiando filas por columnas.

La variable Y toma los valores $y_1, y_2, ..., y_J$ con frecuencias absolutas $n_{,1}, n_{,2}, ..., n_{.J}$ respectivamente.

La frecuencia absoluta para la categoría y_j viene dada por la expresión:

$$n_{.j} = \sum_{i=1}^{I} n_{ij}$$

Y	$n_{.j}$	$f_{.j}$
y_1	$n_{,2}$	$f_{.j}$
y_2	$n_{,2}$	$f_{.j}$
...	...	...
y_j	$n_{.j}$	$f_{.j}$
...	...	...
y_J	$n_{.J}$	$f_{.j}$
Suma	N	1

Tabla 17.10: Frecuencias marginales de la variable Y

Ejemplo 2.

Siguiendo con los datos del Ejemplo 1, vamos a obtener las variables marginales $Curso$ y $Edad$.
La tabla de frecuencias para la variable $Curso$ es:

Curso	$n_{.j}$	$f_{.j}$
Primero	35	0,37
Segundo	28	0,29
Tercero	20	0,21
Cuarto	12	0,13
Suma	95	1

Tabla 17.11: Frecuencias marginales de la variable $Curso$

Y para la variable $Edad$:

Edad	$n_{i.}$	$f_{i.}$
18	20	0,21
19	21	0,22
20	25	0,26
≥ 21	29	0,31
Suma	95	1

Tabla 17.12: Frecuencias marginales de la variable $Edad$

17.4.2 Distribuciones condicionadas

Expresan cómo se distribuye una de las variables sobre un conjunto de individuos que verifican una determinada condición en la otra variable.

Distribución de X condicionada al valor y_j se denota por $X/Y = y_j$, estudia el comportamiento de la variable X sobre aquellos individuos que presentan el valor y_j en la variable Y.
La tabla de frecuencias presenta la siguiente forma:

$X/Y = y_j$	n_i^j	f_i^j
x_1	n_{1j}	f_1^j
x_2	n_{2j}	f_2^j
...	...	...
x_i	n_{ij}	f_i^j
...	...	...
x_I	n_{Ij}	f_I^j
Suma	$n_{.j}$	1

Tabla 17.13: Frecuencias condicionadas de la variable $X/Y = y_j$

En la Tabla 17.13 se observa que la suma de la frecuencia absoluta ahora no es N, si no el número de individuos que verifican la condición, en este caso, $n_{.j}$. Además, $f_i^j = {}^{n_{ij}}/_{n_{.j}}$

Distribución de Y condicionada al valor x_i se denota por $Y/X = x_i$, estudia el comportamiento de la variable Y sobre aquellos individuos que presentan el valor x_i en la variable X.
La tabla de frecuencias presenta la siguiente forma:

$Y/X = x_i$	n_j^i	f_j^i
y_1	n_{i1}	f_1^i
y_2	n_{i2}	f_2^i
...	...	...
y_j	n_{ij}	f_j^i
...	...	...
y_J	n_{iJ}	f_J^i
Suma	$n_{i.}$	1

Tabla 17.14: Frecuencias condicionadas de la variable $Y/X = x_i$

En la Tabla 17.14 se observa que la suma de las frecuencias absolutas es $n_{i.}$ que coincide con los individuos que verifican la condición. La suma de la frecuencia relativa siempre es 1.

Ejemplo 3. Con los datos del Ejemplo 1 se va a calcular la distribución de la variable Edad para los estudiantes de Tercero y la distribución de la variable Curso para los estudiantes de al menos 21 años.

Primero calculamos la variable $Edad$ condicionada a los estudiantes de Tercero

Edad	n_i^3	f_i^3
18	0	0
19	0	0
20	12	0,6
≥ 21	8	0,4
Suma	20	1

Tabla 17.15: Frecuencias condicionadas de la variable $Edad/Curso = Terero$

Como se observa en la Tabla 17.15, el número de estudiantes que verifican la condición, es decir, que el mayor curso en el que están matriculados es Tercero, es 20. El 60 % tienen 20 años y el 40 % al menos 21.

La distribución de la variable Curso condicionada por que tengan al menos 21 años es:

Curso	n_j^4	f_j^4
Primero	4	0,14
Segundo	5	0,17
Tercero	8	0,28
Cuarto	12	0,41
Suma	29	1

Tabla 17.16: Frecuencias condicionadas de la variable $Curso/Edad \geq 21$

Una forma de reconstruir la tabla de frecuencias relativas conjunta es mediante el producto de la distribución marginal por la condicionada para todas las categorías.

$$f_{ij} = f_{.j} f_i^j = f_{i.} f_j^i \quad \forall i, j$$

A partir de los datos del Ejemplo 3 y 4 podemos construir la tabla del Ejemplo 1, así la frecuencia relativa para la categoría (≥ 21, Tercero) se obtiene:

$$f(\geq 21, \text{Tercero}) = f_{43} = f(\geq 21/Tercero) f(Tercero) = f_4^3 f_{,3} = 0,4 * 0,21 = 0,085.$$

De igual forma la frecuencia relativa para la categoría (≥ 21, Segundo) se obtiene:

$$f(\geq 21, \text{Segundo}) = f_{42} = f(Segundo/ \geq 21) f(\geq 21) = f_2^4 f_{2.} = 0,17 * 0,31 = 0,05.$$

17.5 Momentos en las distribuciones bidimensionales

Momentos respecto al origen

Llamamos momento no centrado de orden r y s o momento respecto al origen de orden r y s a la siguiente expresión:

$$a_{rs} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^r y_j^s f_{ij}$$

Entre los momentos más utilizados estan:

- **momentos no centrados de orden 1**

  $a_{10} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^1 y_j^0 f_{ij} = \sum_{i=1}^{I} x_i f_{i.} = \bar{x}$

  $a_{01} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^0 y_j^1 f_{ij} = \sum_{j=1}^{J} y_j f_{.j} = \bar{y}$

- **momentos no centrados de orden 2**

  $a_{20} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^2 y_j^0 f_{ij} = \sum_{i=1}^{I} x_i^2 f_{i.} = a_2(X)$

  $a_{02} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^0 y_j^2 f_{ij} = \sum_{j=1}^{J} y_i^2 f_{.j} = a_2(Y)$

  $a_{11} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i y_j f_{ij} = a_2(Y)$

- **momentos no centrados de orden** $(r, 0)$

  Los momentos no centrados de orden $(r, 0)$ son equivalentes a los momentos no centrados de orden r de la variable X. $a_{r0} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^r y_j^0 f_{ij} = \sum_{i=1}^{I} x_i^r f_{i.} = a_r(X)$

- **momentos no centrados de orden** $(0, s)$

  Los momentos no centrados de orden $(0, s)$ son, igual que en el caso anterior, los momentos no centrados de orden s para la variable unidimensional Y.

  $a_{0s} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_i^0 y_j^s f_{ij} = \sum_{j=1}^{J} y_i^s f_{.j} = a_s(Y)$

Momentos respecto de la media

A los momentos respecto de la media también se les denomina momentos centrados.

Los momentos centrados de orden (r, s) vienen definidos por la siguiente expresión:

$$a_{rs} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^r (y_j - \bar{y})^s f_{ij}$$

Entre los momentos centrados podemos destacar:

- **momentos centrados de orden 1**

  $m_{10} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^1 (y_j - \bar{y})^0 f_{ij} = \sum_{i=1}^{I} (x_i - \bar{x}) f_{i.} = 0$

  $m_{01} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^0 (y_j - \bar{y})^1 f_{ij} = \sum_{j=1}^{J} (y_j - \bar{y}) f_{.j} = 0$

- **momentos centrados de orden 2**

  $m_{20} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^2 (y_j - \bar{y})^0 f_{ij} = \sum_{i=1}^{I} (x_i - \bar{x})^2 f_{i.} = m_2(X) = S^2(X)$

  $m_{02} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^0 (y_j - \bar{y})^2 f_{ij} = \sum_{j=1}^{J} (y_j - \bar{y})^2 f_{.j} = m_2(Y) = S^2(Y)$

  $m_{11} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = S_{XY} = Cov(X, Y)$

- **momentos centrados de orden** $(r, 0)$

  Los momentos centrados de orden $(r, 0)$ son equivalentes a los momentos centrados de orden r de la variable X.

$$m_{r0} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^r (y_j - \bar{y})^0 f_{ij} = \sum_{i=1}^{I} (x_i - \bar{x})^r f_{i.} = m_r(X)$$

- **momentos centrados de orden** $(0, s)$

  Los momentos no centrados de orden $(0, s)$ son, igual que en el caso anterior, los momentos centrados de orden s para la variable unidimensional Y.

  $$m_{0s} = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_i - \bar{x})^0 (y_j - \bar{y})^s f_{ij} = \sum_{j=1}^{J} (y_j - \bar{y})^s f_{.j} = m_s(Y)$$

17.6 Independencia y asociación de las variables

Según hemos indicado en la introducción, uno de los motivos por los que se estudian conjuntamente dos variables es para ver si existe relación entre ellas. En caso de detectar dependencia entre las dos variables, se puede predecir el valor de un individuo en una variable, a partir del valor que tiene en la otra.

Debemos hacer hincapié en que el hecho de que dos variables estén relacionadas no implica que haya una causalidad entre las dos variables.

Dos variables estadísticas son estadísticamente independientes cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra.

Si nos fijamos en la tabla de frecuencias de la variable bidimensional (X, Y),independencia implica que las distribuciones condicionadas no se ven afectadas por la condición, y coinciden en todos los casos con las frecuencias relativas marginales.

$$f_{ij} = f_{i.} f_{.j} \quad \forall i, j$$

Un ejemplo de variables independientes serían las variables X e Y cuyos datos están recogidos en la siguiente tabla de frecuencias:

Valores X \ Y	y_1	y_2
x_1	4	8
x_2	5	10
x_3	8	16
x_4	12	24

Tabla 17.17: Tabla de frecuencias

Calculamos la tabla de frecuencias condicionadas y las marginales

Valores X \Y	y_1	y_2
x_1	0,33	0,67
x_2	0,33	0,67
x_3	0,33	0,67
x_4	0,33	0,67
Marginal Y	0,33	0,67

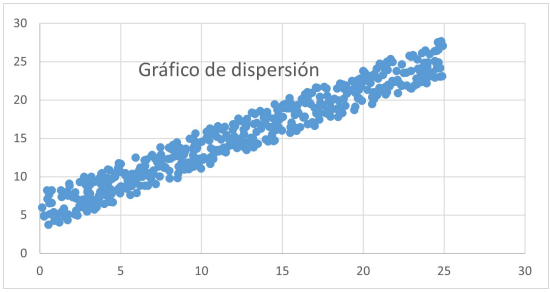
Tabla 17.18: Tabla de distribuciones condicionadas Y/x_i y marginal de Y

Si utilizamos las otras distribuciones condicionadas el resultado sería el mismo.

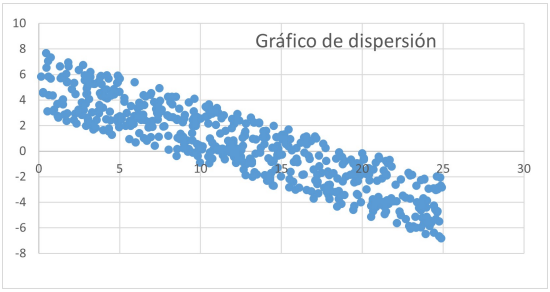
Valores X \Y	y_1	y_2	Marginal de X
x_1	0,14	0,14	0,14
x_2	0,17	0,17	0,17
x_3	0,28	0,28	0,28
x_4	0,41	0,41	0,41

Tabla 17.19: Tabla de distribuciones condicionadas X/y_j y marginal de X

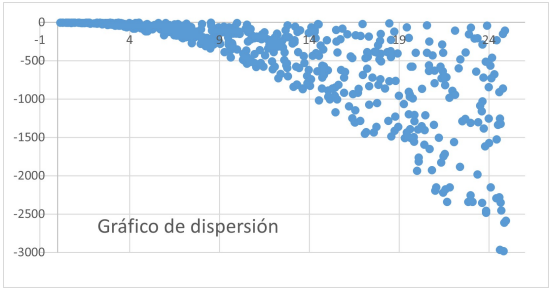
Si las variables no son independientes, una forma de detectar la posible relación entre las variables es gráficamente, mediante el **diagrama de dispersión** o nube de puntos.



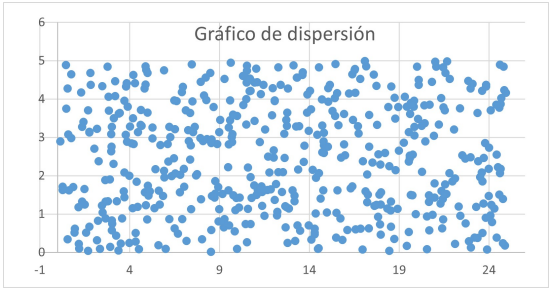
(a) Dependencia lineal positiva



(b) Dependencia lineal inversa



(c) Dependencia no lineal



(d) Independencia

Figura 17.3

Además de gráficamente se pueden valorar distintos coeficientes para determinar el grado de asociación de dos variables estadísticas.

Si al menos una de las variables es cualitativa, es decir, con categorías que no se pueden cuantificar se puede calcular el **coeficiente de contingencia de la** χ^2 a partir de la tabla de frecuencias (si las dos variables son cualitativas a la tabla de fecuencias también se le denomina **tabla de contingencia**).

Se define el coeficiente de contingencia de la χ^2 como:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

siendo $e_{ij} = \frac{n_{i.} n_{.j}}{N}$

Si las variables son independientes la frecuencia observada n_{ij} es igual a la fecuencia esperada e_{ij} para cualquier categoría (i, j) y por lo tanto el valor del coeficiente χ^2 es igual a 0.

El coeficiente también se puede escribir como:

$$\chi^2 = n \left[\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 \right]$$

El mayor inconveniente que tiene este coeficiente es que es proporcional al número de observaciones, y por tanto no tiene una cota superior, por lo que no es muy adecuado su uso.

Para subsanar este problem Karl Pearson definió un nuevo coeficiente basándose en el de la χ^2.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Ahora el índice de Karl Pearson está definido en el intervalo $\left[0, \sqrt{\frac{k-1}{k}}\right]$ siendo $k = \text{mín}\{I, J\}$.

Ejemplo 4.

Queremos saber si los estudiantes del grado de Estadística Aplicada y los del grado de Derecho difieren en sus gustos musicales. Para ello escuchan tres canciones de distintos géneros y deben de elegir la que más les ha gustado. Si los resultados han sido los siguientes:

Estudios \Música	Jazz	Gospel	Rock
Estadística Aplicada	13	9	2
Derecho	2	3	11

Tabla 17.20: Tabla de frecuencias observadas

Primero calculamos la tabla de frecuencias esperadas e_{ij}

Estudios \Música	Jazz	Gospel	Rock
Estadística Aplicada	9	7,2	7,8
Derecho	6	4,8	5,2

Tabla 17.21: Tabla de frecuencias esperadas

Para construir la tabla de frecuencias esperadas hemos calculado las frecuencias de las variables estadísticas marginales y se ha calculado $e_{ij} = \frac{n_{i.}n_{.j}}{N} \quad \forall i, j$.

A partir de las dos tablas de frecuencias calculamos las diferencias $\left(\frac{(n_{ij}-e_{ij})^2}{e_{ij}}\right)$ para calcular el valor del coeficiente de contingencia χ^2

Estudios \Música	Jazz	Gospel	Rock
Estadística Aplicada	1,778	0,450	4,313
Derecho	2,667	0,675	6,469

Tabla 17.22: Tabla de las valores $\left(\frac{(n_{ij}-e_{ij})^2}{e_{ij}}\right)$

Finalmente calculamos el valor del coeficiente $\chi^2 = 16,351$.

Como ya hemos indicado la interpretación de este valor depende del tamaño de la población o de la muestra por lo tanto vamos a calcular el coeficiente $C = \sqrt{\frac{\chi^2}{\chi^2+N}}$ que está definido en $[0, \sqrt{\frac{1}{2}}]$

$$C = \sqrt{\frac{16,351}{16,351+40}} = 0,539$$

Como el valor de C está más cerca del extremo superior que del 0 concluiremos que existe dependencia entre las dos variables.

Si las dos variables son numéricas podemos ver el grado de dependencia lineal a través de la covarianza o del coeficiente de correlación de Pearson.

17.6.1 Covarianza

La covarianza nos indica la relación o dependencia lineal que existe entre las variables estadísticas X e Y. La forma de calcular la covarianza es mediante la media aritmética de los productos de las distancias de la puntuación de cada individuo a su media para cada una de las variables, X e Y. Su expresión matemática es:

$$S_{XY} = \sum_{i,j} f_{i,j}(x_i - \bar{x})(y_j - \bar{y}) = m_{11}$$

Al par $(\bar{x}, \bar{y})$ lo denominamos **Centro de Gravedad**.

Otra forma de calcular la covarianza es a partir del momento no entrado de orden $(1, 1)$

$$S_{XY} = \sum_{i,j} f_{i,j}x_i y_j - \bar{x}\bar{y} = a_{11} - a_{10}a_{10}.$$

Ejemplo 5. Dados 10 valores de una variable bidimensional (X, Y), calcular su covarianza.

x_i	y_i	$x_i y_i$
10	18	182
11	19	204
13	24	309
14	23	332
14	24	343
15	26	380
16	26	412
19	34	632
20	30	578
22	30	658

Tabla 17.23: Valores de las variables X e Y

En este caso para cada i existe un único j, es decir, $f_{i,j} = {}^1\!/_n$. Por lo tanto la fórmula de la covarianza se puede simplificar de la siguiente forma:

$$S_{XY} = \frac{\sum_i x_i y_i}{N} - \bar{x}\bar{y}$$
$$S_{XY} = \frac{4030}{10} - \frac{152}{10}\frac{254}{10} = 15,16$$

La covarianza viene dada en las unidades de la variable X por las unidades de la variable Y.

La covarianza puede tomar cualquier valor real. Si la covarianza está muy próxima a cero, no existe relación entre las variables o si existe, es marcadamente no lineal.

Si es positiva, hay asociación lineal positiva, y si es negativa, hay asociación lineal inversa. Sin embargo, como la covarianza depende de las unidades de medida de las variables, no nos permite cuantificar el grado de asociación lineal.

Un problema que presenta la covarianza es que no se puede comparar la asociación existente entre distintos pares de variables. Para dar solución a este problema se obtiene el **coeficiente de correlación**.

17.6.2 Propiedades de la covarianza

1. Si la covarianza es positiva la dependencia lineal de las dos variables estadísticas es directa.
2. Si la covarianza es negativa, la dependencia lineal es inversa.
3. Si la covarianza es cero diremos que no existe dependencia lineal.
4. Si dos variables son independientes su covarianza es cero pero al contrario no es cierto, si la covarianza es cero no tienen porque ser variables independientes.

17.6.3 Correlación de Pearson

El coeficiente de correlación se va a denotar por la letra griega ρ o por la letra latina r y estudia la relación o dependencia lineal que existe entre las dos variables estadísticas que componen la variable bidimensional.

$$\rho = r = \frac{S_{XY}}{S_X S_Y}$$

La diferencia entre coeficiente de correlación y covarianza es que la covarianza tiene unidades y el coeficiente de correlación no.

Diremos que no existe relación lineal si el valor absoluto del coeficiente de correlación lineal es menor que $0,25$ (en este caso puede existir otro tipo de dependencia, por ejemplo exponencial).

La relación lineal es débil si su valor absoluto está entre $0,25$ y $0,75$ y diremos que es fuerte si es mayor que $0,75$ (los valores aquí mencionados son indicativos, dependiendo del tipo de estudio que estemos realizando podríamos ser más o menos exigentes).

Propiedades del coeficiente de correlación de Pearson

- El signo del coeficiente de correlación viene dado por el signo de la varianza.
- No se debe interpretar el coeficiente de correlación sin haber visto previamente el diagrama de dispersión (podría haber algún dato atípico).

- El valor del coeficiente de correlación lineal está entre el -1 y el 1.

$$\rho \in [-1, 1]$$

- Un coeficiente de correlación alto (en valor absoluto) indica que las observaciones obtenidas de las variables toman valores relacionados entre sí, pero no permite concluir la existencia de ninguna relación de causalidad entre las variables. Por ejemplo, supongamos que se estudian conjuntamente las variables X= Número de mascotas (en una ciudad) e Y =Temperatura media del mes, obteniéndose un coeficiente de correlación de 0,7. Esto no significa que suele haber más mascotas en ciudades con las temperaturas más altas, o que un aumento de mascotas hace que aumente la temperatura de una ciudad, más bien ha sido debido a una casualidad.

Bibliografía

Montiel Torres, AM, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson (página 27).
Peña, D (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial (página 27).

Tema 18

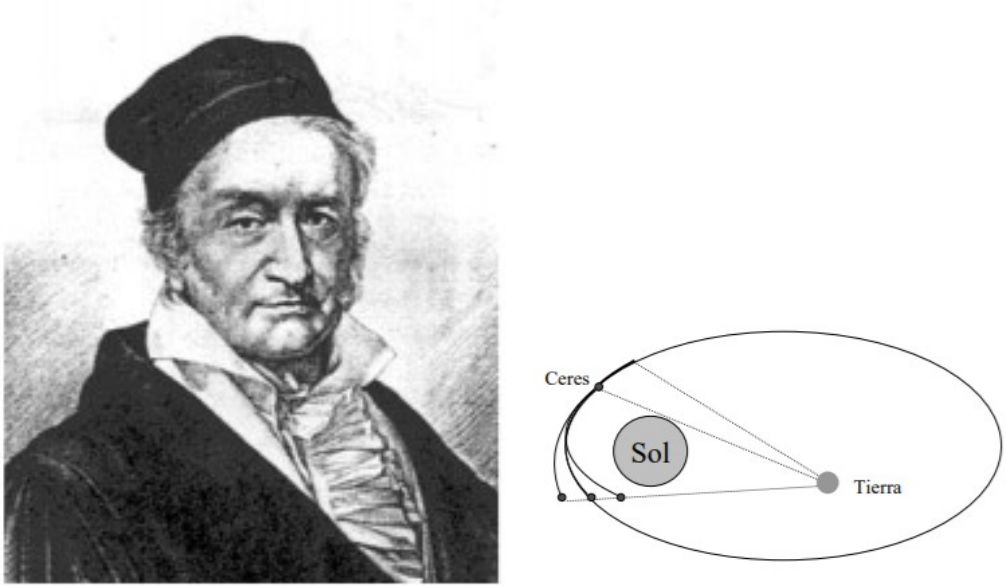
Estadística descriptiva VII. Ajuste por el método de mínimos cuadrados. Varianza residual. Su interpretación.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

AM Montiel Torres, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson
D Peña (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial
Douglas Montgomery, Elizabeth Peck y Geoffrey Vining (2006). *Introducción al análisis de regresión lineal*. México: Limusa Wiley

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.
Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

18.1 Introducción



(a) Carl Friedrich Gauss (b) Planeta Ceres

Figura 18.1: Primera vez que se utilizó el método de mínimos cuadrados

La primera aplicación del método de mínimos cuadrados fue para estimar la trayectoria del planeta enano Ceres, descubierto por el astrónomo italiano Giuseppe Piazzien enero

de 1801. Piazzien fue capaz de seguir su órbita durante 40 días y aunque fueron muchos los astrónomos que utilizaron estos datos para estimar su trayectoria, el único cálculo suficientemente preciso para permitir a Franz Xaver von Zach, astrónomo alemán, reencontrar a Ceres al final del año fue el de Carl Friedrich Gauss, por entonces un joven de 24 años utilizando la técnica de mínimos cuadrados. Este método de mínimos cuadrados no se publicó hasta 1809, y apareció en el segundo volumen de su trabajo sobre mecánica celeste, Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium.

Es un procedimiento de análisis numérico en la que, dados un conjunto de datos (pares ordenados y familia de funciones), se intenta determinar la función continua que mejor se aproxime a los datos (función de regresión). En su forma más simple, busca minimizar la suma de cuadrados de las diferencias ordenadas (residuos) entre los puntos generados por la función y los correspondientes datos.

Este método se utiliza comúnmente para analizar una serie de datos que se obtengan de algún estudio, con el fin de expresar su comportamiento de manera lineal y así minimizar los errores de la data tomada.

18.2 Ajuste por el método de mínimos cuadrados

En esta sección vamos a describir la versión determinista del método de mínimos cuadrados (MCD). Para llevar a cabo este método se necesita la matriz de observaciones y la familia de funciones a las cuales se va a ajustar la nube de puntos. El problema se denomina ajuste de una nube de puntos o regresión bidimensional y consiste en encontrar alguna relación que exprese los valores de una variable en función de los de la otra. La cuestión será elegir la mejor función, y determinar los parámetros (fórmula) de la misma. Esta relación podrá ser utilizada, posteriormente, para hacer predicciones aproximadas; por ejemplo, para hacer previsiones de precios dependiendo de la producción, estimar el volumen de cosecha en función de la lluvia caída, etc.

La matriz de datos está formada por dos columnas correspondientes a la variable estadística bidimensional (X, Y) y tantas filas como individuos observados (N), ((x_1, y_1), (x_2, y_2), ..., (x_N, y_N)). Debemos señalar que la hipótesis de que la variable Y depende de la variable X debe estar fundamentada en algún estudio, el hecho de que exista una función que las relaciona pude deberse a la casualidad y no a una causa-efecto.

En la literatura tenemos ejemplos como que dormir sin quitarse los zapatos está correlacionado con despertarse con dolor de cabeza, obviamente aquí se confunde causalidad y correlación. En muchos de estos casos lo que ocurre es que hay otra variable no obsevada que varía en los individuos muestreados. En este caso puede ser la ingesta de alcohol.

La elección de esa función que mejor defina larelación entre las dos variables es el primer problema que habrá que resolver. En un principio, la observación de la nube de

puntos puede dar una idea de la evolución de los valores de la variable Y en función de los de X.

18.2.1 Ajuste lineal

La familia de funciones que vamos a utilizar es la más sencilla, las funciones lineales:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

El verdadero valor de la variable Y se calcula como la suma de la aproximación de Y (calculada mediante la función lineal $\hat{Y}$) y un error e.

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

siendo:

β_0: la ordenada en el origen, es decir, el valor de la variable dependiente Y cuando el predictor o variable independiente es cero.

β_1: la variación media de la variable Y cuando se incrementa en una unidad de la variable predictora X. Se conocen como coeficiente de regresión de Y sobre X.

$\hat{y}_i$: la estimación de la variable Y para el individuo i-ésimo a través del modelo de regresión.

e_i: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo para cada individuo.

El método de mínimos cuadrados lo que hace es estimar los parámetros β_0 y β_1 para minimizar la discrepancia entre los datos observados de la variable Y y los estimados por la función $\hat{Y}$.

$$\text{mín} f(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2$$

Es importante destacar que el MCD es un método de aproximación pues no utiliza hipótesis probabilísticas sobre los datos. Para calcular el mínimo de la función $f(\beta_0, \beta_1)$ derivamos respecto de los parámetros e igualamos a 0.

$$\frac{\partial \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2}{\partial \beta_0} = 0$$

$$\frac{\partial \sum_{i=1}^{N}(y_i-(\beta_0+\beta_1 x_i))^2}{\partial \beta_1}=0$$

obteniendo el siguiente sistema de ecuaciones:

$$\begin{aligned}\tfrac{\partial}{\partial \beta_0}\textstyle\sum_{i=1}^{N}(y_i-(\beta_0+\beta_1 x_i))^2 &= -2\textstyle\sum_{i=1}^{N}(y_i-(\beta_0+\beta_1 x_i))=0\\ \tfrac{\partial}{\partial \beta_1}\textstyle\sum_{i=1}^{N}(y_i-(\beta_0+\beta_1 x_i))^2 &= -2\textstyle\sum_{i=1}^{N}(y_i-(\beta_0+\beta_1 x_i))x_i=0\end{aligned} \tag{18.1}$$

desarrollando las ecuaciones llegamos a las siguientes expresiones:

$$\begin{aligned}\textstyle\sum_{i=1}^{N} y_i &= N\beta_0+\beta_1\textstyle\sum_{i=1}^{N} x_i\\ \textstyle\sum_{i=1}^{N} y_i x_i &= \beta_0\textstyle\sum_{i=1}^{N} x_i+\beta_1\textstyle\sum_{i=1}^{N} x_i^2\end{aligned} \tag{18.2}$$

de la primera ecuación despejamos β_0 obteniendo

$$\beta_0=\bar{y}-\beta_1\bar{x} \tag{18.3}$$

sustituyendo el valor de β_0 en la segunda ecuación podemos despejar β_1

$$\begin{aligned}0= &\ \textstyle\sum_{i=1}^{N} y_i x_i-(\bar{y}-\beta_1\bar{x})\textstyle\sum_{i=1}^{N} x_i-\beta_1\textstyle\sum_{i=1}^{N} x_i^2=\\ &=(\textstyle\sum_{i=1}^{N} y_i x_i-N\bar{y}\bar{x})-\beta_1(\textstyle\sum_{i=1}^{N} x_i^2-N\bar{x}^2)= \qquad (18.4)\\ &=S_{XY}-\beta_1 S_X^2\end{aligned}$$

$$\beta_1=\frac{S_{XY}}{S_X^2} \tag{18.5}$$

Por lo tanto la función que minimiza los errores al cuadrado se denomina **recta de regresión de** Y **sobre** X y viene dada por la expresión:

$$\hat{Y}=\bar{y}-\frac{S_{XY}}{S_X^2}\bar{x}+\frac{S_{XY}}{S_X^2}X \tag{18.6}$$

otra forma de expresarla es:

$$\hat{Y}-\bar{y}=\frac{S_{XY}}{S_X^2}(X-\bar{x}) \tag{18.7}$$

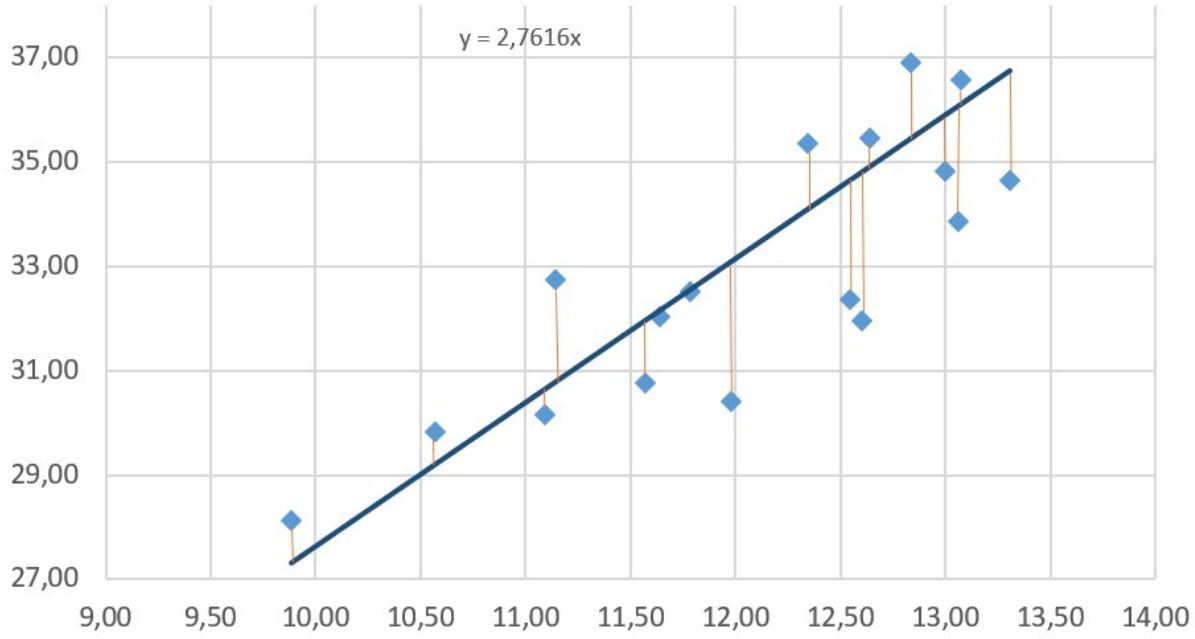


Figura 18.2: Ajuste de una nube de puntos

En la Figura 18.2 está representada la nube de puntos y la recta de regresión de Y sobre X que mejor ajusta los datos ($\hat{Y} = 2,761X$).

Si se analizan los residuos ($e_i = y_i - \hat{y}_i$) esta variable estadística debe tener una media aritmética igual a 0. Además, si representamos el gráfico de dispersión de las variables (X, e), representando en el eje de abscisas a la variable X y en el eje de ordenadas a la variable e los valores de los errores deben ser aleatorios.

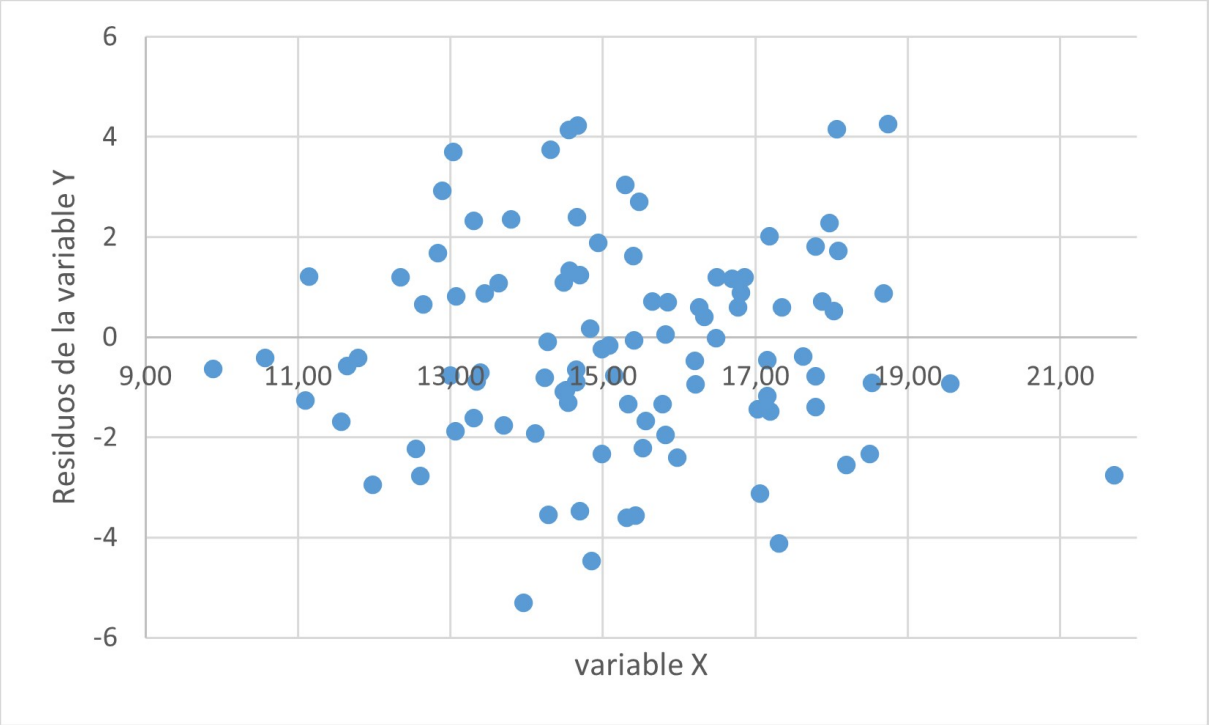


Figura 18.3: Residuos de la variable Y respecto a los valores de la variable X

Ejemplo 1. Con los datos de la Tabla 18.1 se va a construir la mejor recta de regresión de Y sobre X utilizando la técnica de mínimos cuadrados.

X	9,00	10,00	11,00	12,00	13,00	14,00	15,00
Y	29,13	31,01	32,24	37,21	36,81	40,42	43,08

Tabla 18.1: Valores de las variables X e Y

Para construir la recta de regresión se necesitan conocer la covariaza entre X e Y (S_{XY}), el vector de medias $(\bar{x}, \bar{y})$ y la varianza de X (S_X^2).

X	Y	X^2	$X*Y$
9,00	29,13	81,00	262,18
10,00	31,01	100,00	310,18
11,00	32,24	121,00	354,69
12,00	37,21	144,00	446,59
13,00	36,81	169,00	478,63
14,00	40,42	196,00	566,00
15,00	43,08	225,00	646,24
SUMA			
84,00	249,94	1036,00	3064,53

Tabla 18.2: Cálculos intermedios

A partir de los datos de la Tabla 18.2 obtenemos:

vector de medias:

$$(\bar{x}, \bar{y}) = (\frac{84}{7}, \frac{249,94}{7}) = (12; 35,71)$$

covarianza de X e Y:

$$S_{XY} = \frac{3064,53}{7} - 12 * 35,71 = 9,32$$

varianza de X:

$$S_X^2 = \frac{1036}{7} - 12^2 = 4$$

Sabiendo que la recta de regresión de Y sobre X es:

$$\hat{Y} - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

La recta de regresión obtenida es:

$$\hat{Y} - 35,71 = 2,33(X - 12)$$

Por lo tanto la pendiente de la recta es $\beta_1 = 2,33$. Lo que indica que por cada unidad que se modifique la variable independiente X, la variable dependiente Y variará en $2,33$ unidades.

Si la variable X toma el valor 0, en ese caso la variable Y tomará el valor $\beta_0 = \bar{y} - \beta_1 * \bar{x} = 7,74$. Este dato no siempre tiene significado pues el valor cero puede no existir para la variable independiente X.

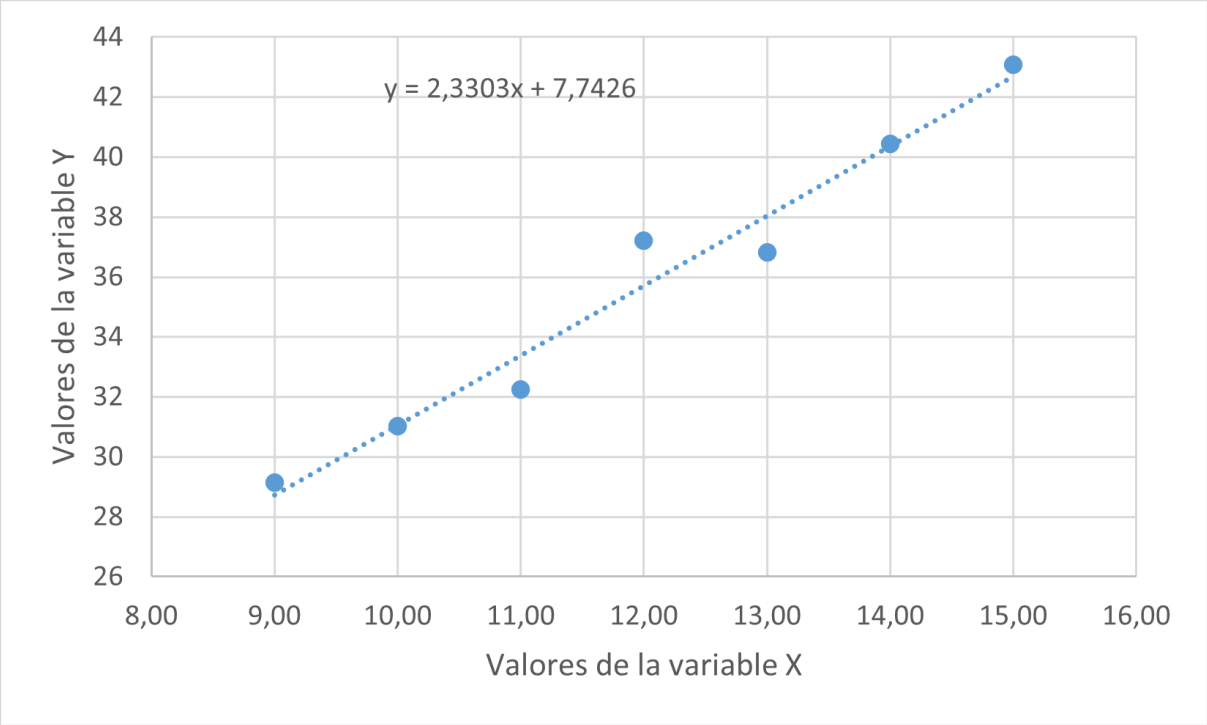


Figura 18.4: Ajuste de la nube de puntos a la recta de regresión

18.2.2 Transformaciones para conseguir linealidad

Aunque hemos desarrollado el caso de funciones lineales puede ocurrir que la curva que mejor ajuste la nube de puntos, no sea una recta. En datos que tienen un crecimiento o decrecimiento constante, la mejor familia de funciones para ajustar la nube de puntos por el MCD es:

$$\hat{Y} = \beta_0 e^{\beta_1 X}$$

A esta regresión se le denomina **regresión exponencial** y se puede transformar en una regresión lineal mediante una transformación logaritmica de los datos.

$$Ln(\hat{Y}) = Ln(\beta_0 e^{\beta_1 X}) = Ln(\beta_0) + \beta_1 X$$

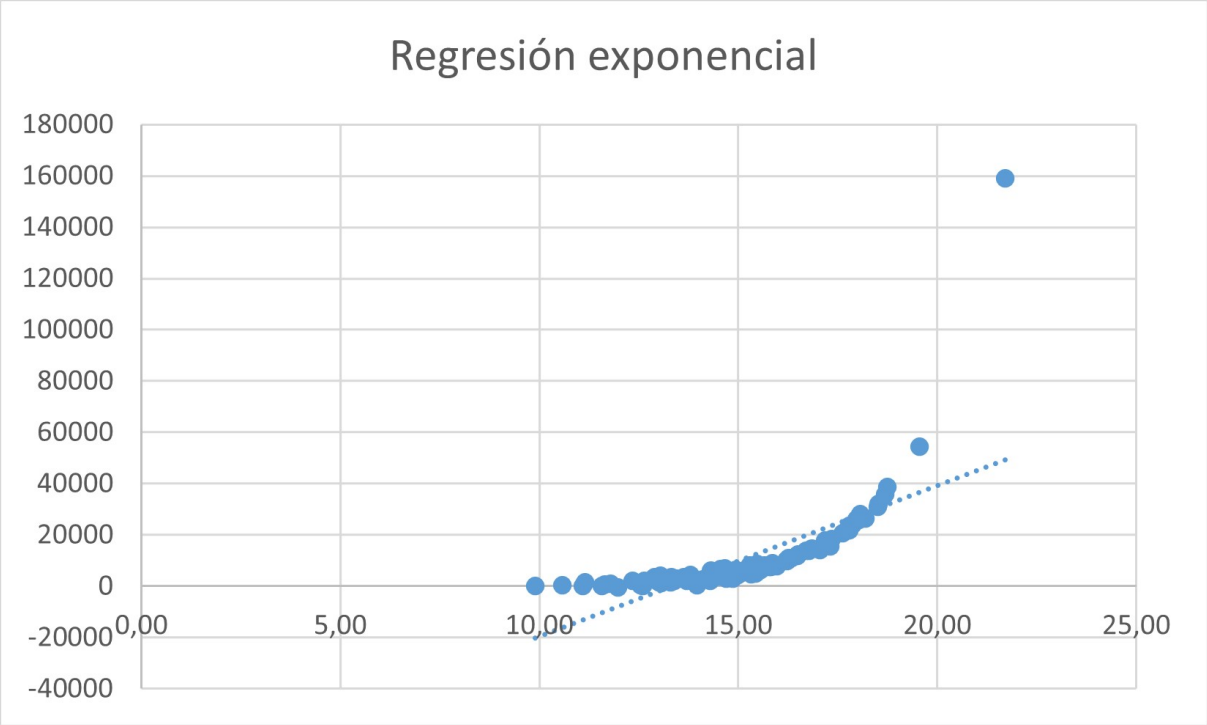


Figura 18.5: Regresión exponencial

Si observamos la Figura 18.5 se comprueba que la función que mejor ajusta la nube de puntos no es la lineal. Si se desea realizar predicciones mediante el ajuste lineal se comprueba que los errores cada vez son mayores, es decir la variable de los residuos no es aleatoria si no que va aumentando en valor absoluto a medida que nos separamos de la media de X. En este caso no se debe trabajar con los datos originales si no con $Y' = Ln(Y)$ y X.

Otro ejemplo de ajuste no lineal es **la regresión potencial** que viene expresada por:

$$\hat{Y} = \beta_0 X^{\beta_1}$$

Igual que en el caso de la regresión exponencial, podemos transformar los datos para poder realizar una regresión lineal con los datos transformados.

$$Ln(\hat{Y}) = Ln(\beta_0 X^{\beta_1}) = Ln(\beta_0) + \beta_1 Ln(X)$$

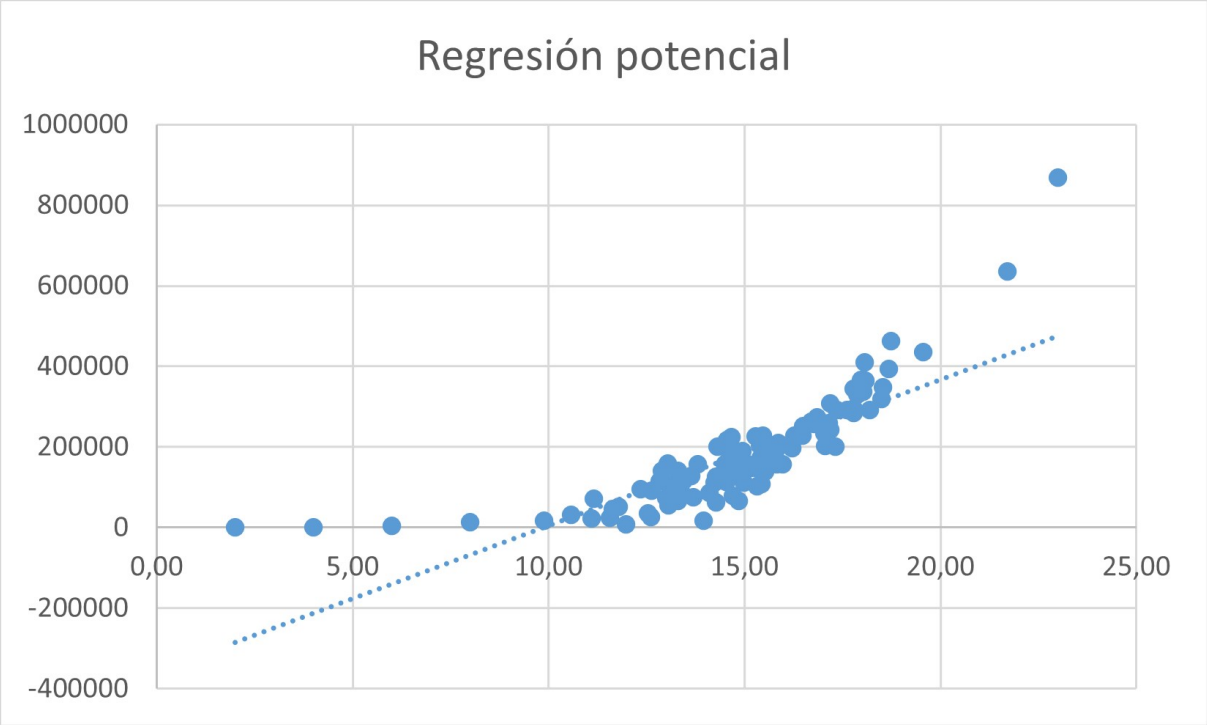


Figura 18.6: Regresión potencial

En la Figura 18.6 se puede observar como la recta de regresión pasa por el centro de gravedad de los datos pero a medida que nos acercamos a los valores extremos el error que cometemos es más grande. En este caso no se debe trabajar con los datos originales si no con $Y' = Ln(Y)$ y $X' = Ln(X)$.

18.2.3 regresión polinomial

La dependencia entre la variable dependiente y la regresora frecuentemente no es lineal. No obstante, se ajusta al modelo lineal simple a no ser que el modelo no lineal se demuestre significativamente superior al lineal.

La familia de funciones que definen la regresión polinomial viene dada por:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k$$

Veamos el desarrollo del cálculo de los parámetros β_0, β_1 y β_2 por el método de los mínimos cuadrados (MCD) para el caso de polinomios de grado 2.

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Definimos los residuos como:

$$e_i = y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2$$

Si utilizamos la técnica de MCD para calcular los parámetros β_0, β_1 y β_2 se debe minimizar la suma de los residuos

$$\text{mín} f(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2$$

por lo tanto, las derivas parciales respecto de cada uno de los parámetros deben ser igual a cero.

$$\frac{\partial \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}{\partial \beta_0} = 0$$

$$\frac{\partial \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}{\partial \beta_1} = 0$$

$$\frac{\partial \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2}{\partial \beta_2} = 0$$

es decir:

$$\begin{aligned}
&\tfrac{\partial}{\partial \beta_0} \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2 = -2 \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)) = 0 \\
&\tfrac{\partial}{\partial \beta_1} \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2 = -2 \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)) x_i = 0 \\
&\tfrac{\partial}{\partial \beta_2} \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2 = -2 \textstyle\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)) x_i^2 = 0
\end{aligned} \tag{18.8}$$

obteniendo las siguientes ecuaciones:

$$\begin{aligned}
\textstyle\sum_{i=1}^{N} y_i &= N\beta_0 + \beta_1 \textstyle\sum_{i=1}^{N} x_i + \beta_2 \textstyle\sum_{i=1}^{N} x_i^2 \\
\textstyle\sum_{i=1}^{N} y_i x_i &= \beta_0 \textstyle\sum_{i=1}^{N} x_i + \beta_1 \textstyle\sum_{i=1}^{N} x_i^2 + \beta_2 \textstyle\sum_{i=1}^{N} x_i^3 \\
\textstyle\sum_{i=1}^{N} y_i x_i^2 &= \beta_0 \textstyle\sum_{i=1}^{N} x_i^2 + \beta_1 \textstyle\sum_{i=1}^{N} x_i^3 + \beta_2 \textstyle\sum_{i=1}^{N} x_i^4
\end{aligned} \tag{18.9}$$

La solución del sistema nos genera la parábola que mejor se ajusta a la nube de puntos.

Ejemplo 2.
Veamos un ejemplo de regresión polinomial de grado 2 (**regresión parabólica o cuadrática**) para la resolución se utilizará el sotfware IBM SPSS Statistics 27.

Dada la siguiente matriz de datos:

X	9,89	10,57	12,35	13,06	14,27	16,48	17,35	18,52	23,00
Y	-404,11	-462,43	-632,10	-711,37	-848,94	-1134,05	-1256,86	-1435,21	-2210,54

Tabla 18.3: Valores de los datos de las variables X e Y

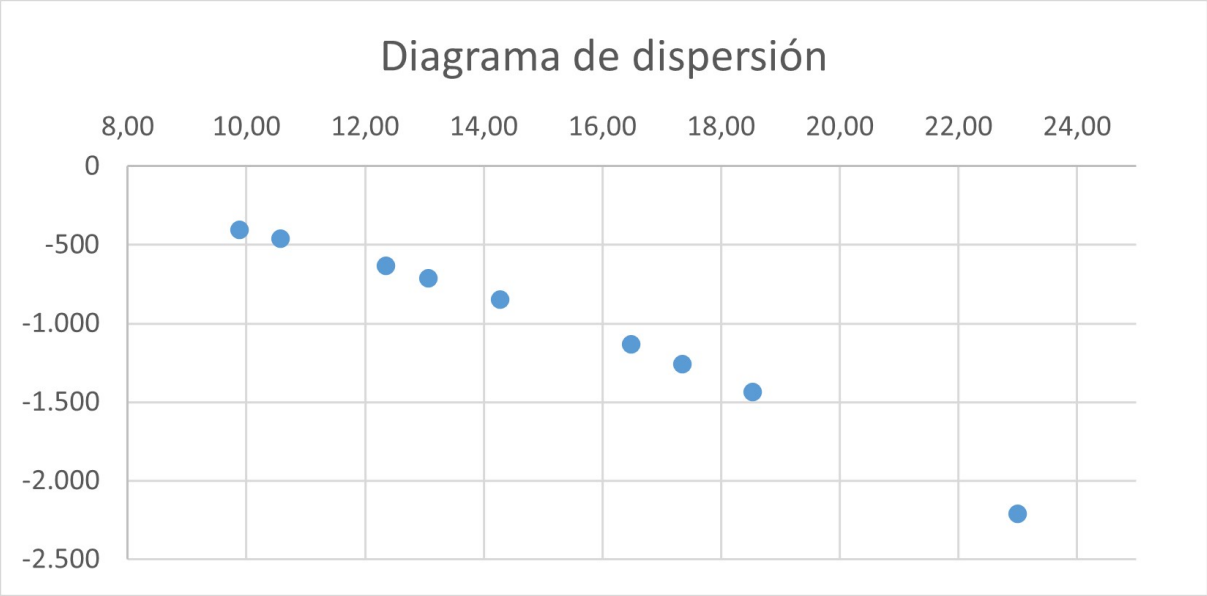


Figura 18.7: Diagrama de dispersión para las variables X e Y

Construimos la tabla de los cálculos intermedios para poder resolver el sistema de ecuaciones 18.9

X	Y	X^2	X^3	X^4	$Y * X$	$Y * X^2$
9,89	-404,1179093	97,73	966,22	9552,10	-3995,15	1614510,579
10,57	-462,4365596	111,67	1180,08	12470,41	-4886,78	2259823,573
12,35	-632,1045657	152,45	1882,34	23241,57	-7804,68	4933376,268
13,06	-711,378046	170,60	2228,18	29102,66	-9291,45	6609735,658
14,27	-848,9424187	203,77	2908,84	41523,32	-12118,57	10287965,48
16,48	-1134,053835	271,68	4477,93	73807,97	-18692,16	21197915,67
17,35	-1256,866985	301,06	5223,83	90639,77	-21808,16	27409956,14
18,52	-1435,210761	343,17	6357,08	117763,30	-26586,93	38157846,45
23,00	-2210,545202	529,00	12167,00	279841,00	-50842,54	112389732,1

Tabla 18.4: Valores intermedios para plantear el sistema de ecuaciones 18.9

Para la construcción de la parábola que mejor ajusta la nube de puntos utilizamos el sotfware IBM SPSS Statistics 27, obteniendo el siguiente resultado:

Variable dependiente: Y

Ecuación	Resumen del modelo					Estimaciones de parámetro		
	R cuadrado	F	df1	df2	Sig.	Constante	b1	b2
Lineal	,984	424,349	1	7	,000	1025,106	-135,225	
Cuadrático	1,000	726961,665	2	6	,000	19,199	-1,826	-4,136

La variable independiente es X.

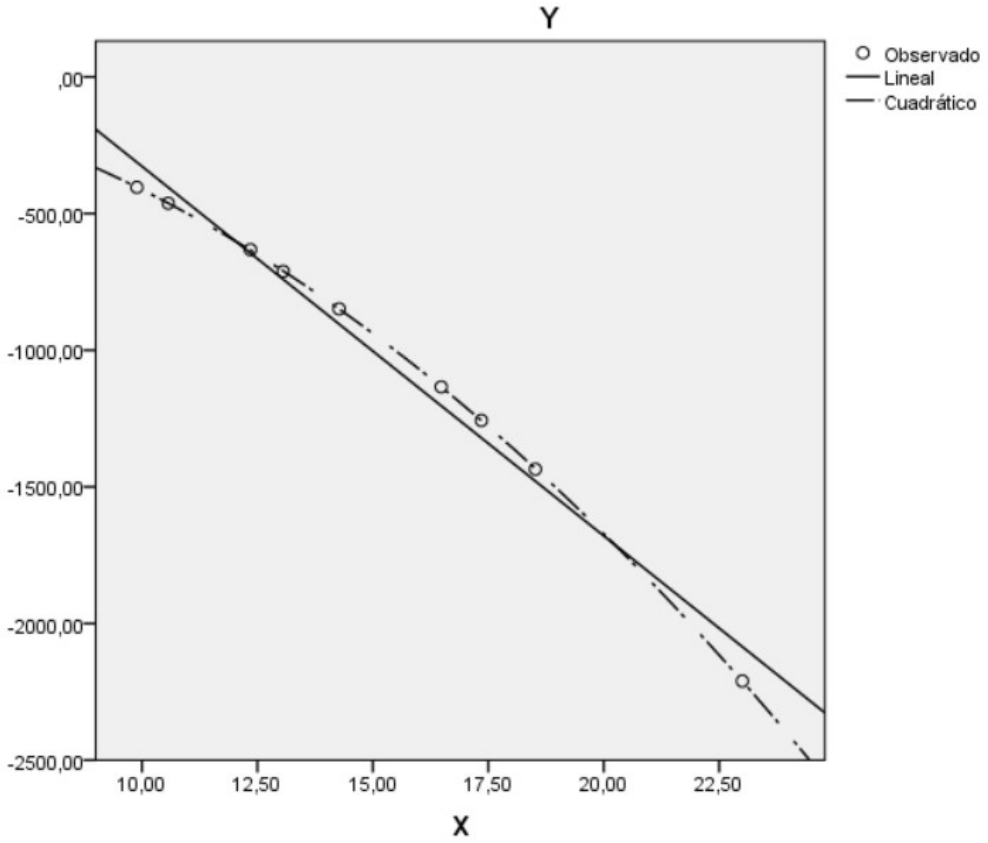


Figura 18.8: Regresión cuadrática o parabólica

en la Figura 18.8 se pueden observar las estimaciones de los parámetros de la parábola: $\beta_0 = 19,199$, $\beta_1 = -1,826$ y $\beta_2 = -4,136$.

Por lo tanto, la función polinomial de grado 2 que mejor ajusta la nube de puntos es:

$$\hat{y}_i = 19,199 - 1,826x_i - 4,136x_i^2$$

.

Si queremos simplificar el modelo de regresión, podemos ajustar los datos mediante un modelo lineal, en ese caso los valores de los parámetros son: $\beta_0 = 1025,106$ y $\beta_1 = -135,225$.

La recta de regresión obtenida es:

$$\hat{y}_i = 1025,106 - 135,225x_i$$

.

En el gráfico de la Figura 18.8 podemos ver la diferencia entre ajustar por un modelo lineal o cuadrático

18.3 Varianza residual

Hemos visto en la sección anterior que el criterio de mínimos cuadrados utiliza la media de los cuadrados de los residuos como medida del error que se comete, cuando ajustamos el valor de la variable Y por su estimación $\hat{Y}$, esta media recibe el nombre de **Varianza Residual (S_e^2)**.

La Varianza Residual se utiliza como medida de la bondad del ajuste. Cuanto menor sea la Varianza Residual, menores serán los residuos y por lo tanto mejor será el ajuste de la curva a la nube de puntos. Sin embargo, recordemos que las varianzas están medidas en las mismas unidades de la variable Y pero al cuadrado. Por lo tanto, ¿a partir de que valores dicha varianza es suficientemente pequeña para considerar que el ajuste realizado es un buen ajuste? Ese problema se soluciona quitando unidades mediante la definición de un indicador o coeficiente.

La varianza de Y se puede expresar como suma de dos varianzas parciales:

$$S_Y^2 = \underbrace{\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N}}_{VarianzadeY} = \underbrace{\frac{\sum_{i=1}^{N}(y_i-\hat{y})^2}{N}}_{VarianzaResidual} + \underbrace{\frac{\sum_{i=1}^{N}(\hat{y_i}-\bar{y})^2}{N}}_{VarianzaExplicada}$$

Por un lado la **Varianza Residual** o cantidad de varianza que no se ha podido explicar mediante el modelo de regresión

$$S_e^2 = \frac{\sum_{i=1}^{N}(y_i-\hat{y})^2}{N}$$

y por otro la **Varianza Explicada** mediante la aproximación de los valores de Y por sus estimaciones a través del modelo de regresión

$$S_{\hat{Y}}^2 = \frac{\sum_{i=1}^{N}(\hat{y_i}-\bar{y})^2}{N}$$

obteniendo:

$$S_Y^2 = S_e^2 + S_{\hat{Y}}^2$$

Ejemplo 3.

Con los datos del Ejemplo 1 veamos como se calcularía la varianza residual.

X	Y	$\hat{Y}$	e_i	$\hat{Y}^2$	e_i^2
9,00	29,13	28,72	0,42	824,57	0,17
10,00	31,01	31,05	-0,03	963,83	0,00
11,00	32,24	33,38	-1,13	1113,95	1,28
12,00	37,21	35,71	1,51	1274,93	2,28
13,00	36,81	38,04	-1,22	1446,78	1,48
14,00	40,42	40,37	0,06	1629,48	0,00
15,00	43,08	42,70	0,39	1823,04	0,15
SUMA					
84,00	249,94	249,94	0,00	9076,58	5,37

Tabla 18.5: Valores intermedios para calcular las varianzas

A partir de los datos de la Tabla 18.5 podemos deglosar la varianzas de Y en la varianza residual y la varianza explicada por la recta de regresión.

$$S_e^2 = \frac{5,37}{7} = 0,77$$

$$S_{\hat{Y}}^2 = \frac{9{,}076,58}{7} - 35,71^2 = 21,72$$

Puesto que:

$$S_Y^2 = S_e^2 + S_{\hat{Y}}^2$$
$$S_Y^2 = 21,72 + 0,77 = 22,49$$

18.3.1 Coeficiente de Determinación:

La definición del coeficiente de determinación se basa en la descomposición de la varianza de Y.

Definimos el **coeficiente de determinción** y lo denotamos por R^2 como la proporción de varianza explicada por el modelo de regresión.

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} \tag{18.10}$$

Si lo queremos expresar en función de la varianza residual nos queda:

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} = \frac{S^2_Y - S^2_e}{S^2_Y} = 1 - \frac{S^2_e}{S^2_Y} \tag{18.11}$$

Al definir el R^2 como el cociente de dos varianzas de la variable Y, este ya no tiene unidades y por lo tanto podemos analizar, a partir de su valor, la bondad de ajuste de cualquier modelo de regresión.

Si se observa la ecuación 18.10, R^2 se puede expresar como $1 - \epsilon$, siendo $\epsilon > 0$. Esto nos indica que R^2 está definido en el intervalo $[0, 1]$.

Ejemplo 4. siguiendo con los datos del Ejemplo 3, veamos la bondad de ajuste de la recta de regresión de Y sobre X.

Si utilizamos la ecuación 18.10:

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} = \frac{21,72}{22,49} = 0,966$$

lo que significa que el $96,6\,\%$ de la varianza de Y está explicada por la relación lineal que existe entre las dos variables (recta de regresión).

Relación entre el coeficiente de correlación y el coeficiente de determinación en el caso de regresión lineal

Recordemos que el coeficiente de correlación lineal mide la relación o dependencia lineal que existe entre las dos variables estadísticas que componen la variable bidimensional.

$$\rho = r = \frac{S_{XY}}{S_X S_Y}$$

Para el caso de la regresión lineal tenemos:

$$\hat{y}_i = \beta_0 + \beta_1 x_i = \beta_0 + \frac{S_{XY}}{S^2_X} x_i$$

Aplicando las propiedades de la varianza para transformaciones lineales se obtiene:

$$S^2_{\hat{Y}} = (\frac{S_{XY}}{S^2_X})^2 * S^2_X = \frac{S^2_{XY}}{S^2_X}$$

a continuación dividimos por S^2_Y para expresar R^2 en función del coeficiente de correlación lineal r:

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} = \frac{S^2_{XY}}{S^2_Y S^2_X} = \rho^2$$

18.3.2 Interpretación del Coeficiente de Determinación

Puesto que el Coeficiente de Determinación no tiene unidades podemos interpretar su valor con independencia de la variable bidimensional (X, Y) que lo haya generado.

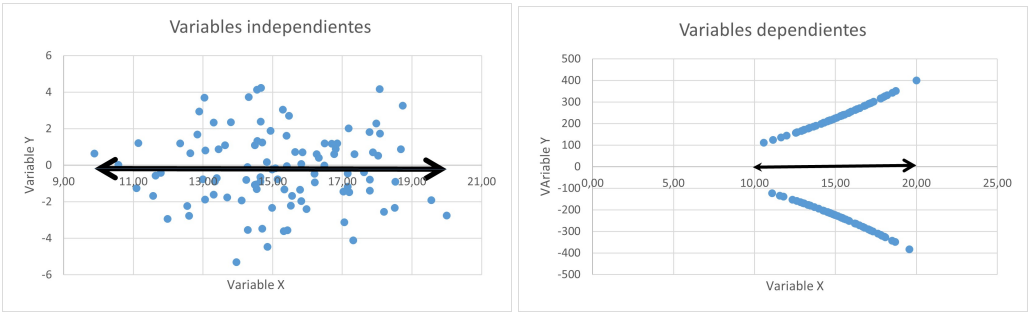
Al estar definido entre 0 y 1 podemos interpretar a R^2 como el tanto por uno de la Varianza de Y explicada por el modelo de regresión.

1. Si $R^2 = 0$, como

   $R^2 = 1 - \frac{S^2_e}{S^2_Y}$

   La varianza de Y coincide con la varianza residual y podemos concluir que la varianza explicada es cero. Es decir, el modelo de regresión no explica las variaciones de la variable Y, es el peor ajuste que se puede realizar mediante las técnicas de mínimos cuadrados.

   En este caso la recta de regresión es paralela al eje donde está representada la variable independiente X y eso es debido a que todas las distribuciones de Y condicionada por X_i tienen la misma media.



(a) Variables independientes (b) Dependencia no lineal

Figura 18.9: Posibles relaciones entre variables con $R^2 = 0$

Si comparamos los dos gráficos de la Figura 18.9 se observa que no es equivalente que R^2 sea 0 a que las variables sean independientes. En el caso (a) las dos variables son independientes mientras que en el casoo (b) existe una dependencia pero no lineal.

2. Si $R^2 = 1$, como

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y}$$

La varianza de Y queda totalmente explicada por el modelo de regresión obtenido mediante la técnica de mínimos cuadrados. Todas las estiamciones coinciden con los valores de la variable dependiente (Y) esto implica que Y tiene una **dependencia funcional** de X.

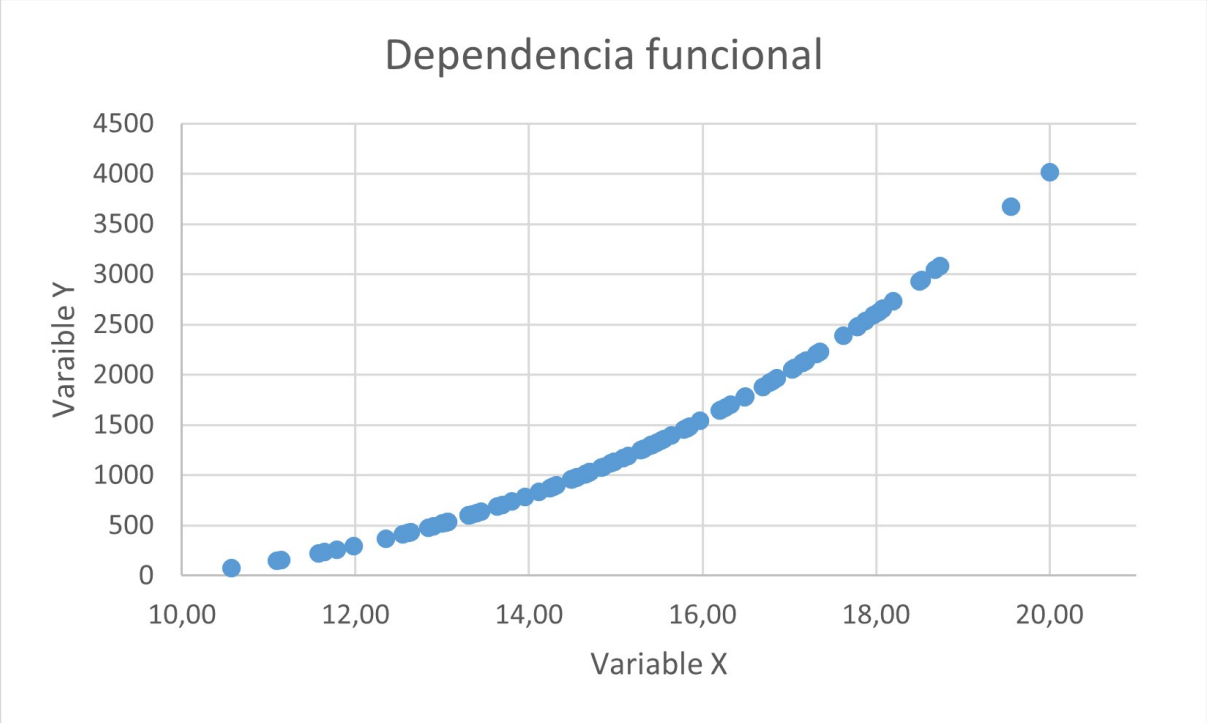
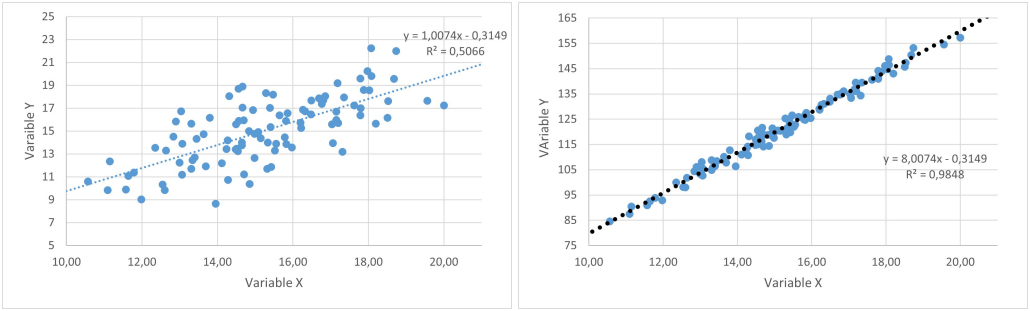


Figura 18.10: Dependencia funcional entre dos variables $y_i = x_i - 10x_i^2 + x_i^3$

La dependencia funcional es un caso extremo que rara vez se encuentra en la práctica. Un fenomeno que se rige por una dependencia funcional es un fenómeno determinista pues conociendo el valor de la variable X, la variable Y queda totalmente determinada. Un ejemplo de dependencia funcional es la velocidad y el espacio, sabiendo a la velocidad a la que nos desplazamos podemos calcular el espacio que vamos a recorrer en una unidad de tiempo.

3. Si $0 \leq R^2 \leq 1$, existe una dependencia estadística.

Si R^2 se aproxima al 0 diremos que el ajuste es muy malo y cuanto más se aproxime a 1 mejor será el modelo de regresión.



(a) Dependencia lineal débil (b) Dependencia lineal muy fuerte

Figura 18.11: Representaciones para distintos valores de R^2

Como se puede observar en la Figura 18.11 a medida que el coeficiente de determinación (R^2) se aproxima a 1 la nuebe de puntos se ajusta más a la recta de regresión y por lo tanto los e_i son más pequeños.

El modelo (a) de regresión de la Figura 18.11 indica que la recta de regresión solamente explica el $50,66\,\%$ de la varianza de la variable Y mientras que en el modelo (b) explica un $98,48\,\%$.

Hay varias razones principales por las que los valores bajos del R^2 podrían considerarse adecuados.

En algunos campos, se espera que los valores del R^2 sean bajos. Por ejemplo, cualquier disciplina que intenta predecir el comportamiento humano, como la psicología, normalmente tiene valores del R^2 inferiores al $50\,\%$. Los seres humanos son simplemente más difíciles de predecir que, por ejemplo, los procesos físicos.

Además, si el valor del R^2 es bajo pero se tiene predictores estadísticamente significativos, se pueden obtener conclusiones importantes acerca de la asociación entre los cambios en los valores de los predictores y los cambios en el valor de la variable dependiente.

Un R^2 bajo es más problemático cuando se desea crear predicciones que sean muy precisas.

Por el contrario, un R^2 alto no necesariamente indica que el modelo tiene un buen ajuste. Esto podría sorprendernos, pero para validar un modelo, además del R^2 se debe examinar la gráfica de los residuos, esta debe ser aleatoria y homocedástica.

Bibliografía

Montiel Torres, AM, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson (página 45).
Peña, D (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial (página 45).
Montgomery, Douglas, Elizabeth Peck y Geoffrey Vining (2006). *Introducción al análisis de regresión lineal*. México: Limusa Wiley (página 45).

Tema 19

Estadística descriptiva VIII. Recta de regresión. Coeficiente de correlación lineal y cálculo del mismo. Posiciones de las rectas de regresión según el valor del coeficiente de correlación.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

AM Montiel Torres, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson
D Peña (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial
Douglas Montgomery, Elizabeth Peck y Geoffrey Vining (2006). *Introducción al análisis de regresión lineal*. México: Limusa Wiley Carlos Camacho, AM López y MA Arias (2006). "Regresión lineal simple". En: *de Apuntes no publicados de la asignatura Análisis de datos II de la licenciatura de Psicología, Universidad de Sevilla*

19.1 Introducción

En el tema 17 se han desarrollado conceptos como el de la distribución condicionada o el de la correlación entre dos variables. A través de estos conceptos se han podido detectar que en algunas ocasiones existen relaciones entre dos variables. De forma que la información de una nos ayuda a predecir el valor de la otra.

Día a día nos encontramos fenómemos que nos gustaría predecir a través de modelos estadísticos. Sin embargo, no sabemos cual es la causa principal que los origina. En regresión debemos tener cuidado con la interpretación del coeficiente de correlación lineal. La ausencia de correlación no implica la independencia de las variables y al contrario puede ser que exista una correlación lineal próxima a uno y sin embargo sea por **casualidad** y no **la causa**. Cuando formulamos un modelo de regresión nos debemos de basar en una teoría que sustente la relación entre las dos variables.

Supongamos que se desea saber la nota media de la EvAU en la Comunidad de Madrid a partir de la nota media de segundo de bachiller. Parece lógico ver cuales han sido las notas de los estudiantes que han tenido la misma media en segundo de bachiller y calcular la media de esta distribución condicionada. Este procedimiento tiene el inconveniente de que si la variable independiente es continua no podemos calcular todas las

distribuciones condicionadas.

Para resolver este problema lo que se suele hacer es modelizar, es decir buscar una función que relacione las dos variables. Así, en el caso de las notas de la EvAU dibujaríamos el diagrama de dispersión poniendo en el eje de abscisas las nota media de segundo de bachiller y en el eje de ordenadas la nota media de la EvAU y analizaríamos si una recta es la función que más ajusta la nube de puntos.

En este ejemplo parece sencillo identificar cual es la variable independiente y cual la variable dependiente, pero no siempre es así. El primer problema con el que se encuentra el investigador es como establecer de forma clara las variables que son causa (variables independientes), las que son efecto (variable dependiente) y las que interviniendo en el experimento que deben ser controladas para no intruducir ruido en el análisis de datos. La variable independiente debe ser controlada y medida con instrumentos apropiados para poder cuantificar los efectos que se producen en la variable dependiente debido a los cambios producidos en la variable independiente.

19.2 Recta de regresión

Figura 19.1

La teoría de la regresión tiene por objeto definir la estructura de dependencia que mejor explique el comportamiento de una variable (dependiente o explicada) en función de otra (independiente o explicativa).

En el caso de la recta de regresión el modelo matemático que mejor explica la relación entre las dos variables es una recta.

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

A la variable dependiente habitualmente se le denomina Y y a la variable independiente X.

19.2.1 Hipótesis y estimación

Para que un modelo de regresión lineal sea bueno, no solo nos debemos fijar en la media de los errores que se comenten, cuando le utilizamos para predecir la variable dependiente.

Para que los resultados de la regresión lineal sean fiables se necesita tener al menos 30 datos y así poder utilizar el Teorema Central del Límite y afirmar que las estimaciones son consistentes. Además, las variables deben verificar las siguientes hipótesis:

1. La relación entre las variables debe ser lineal o existe una transformación en los datos después de la cual se consigue la linealidad.
2. Los residuos o perturbaciones deben tener media cero, homocedasticidad y no ser autocorreladas (esfericidad de los residuos).

19.2.2 Limitaciones del método de mínimos cuadrados

Tal y como se ha visto en el tema 18, el método de regresión por mínimos cuadrados es muy útil para estudiar una variable respuesta continua en función del predictor o variable independiente. Sin embargo, bajo determinadas circunstancias y a pesar de existir una relación lineal entre las variables, el método de mínimos cuadrados puede no ser adecuado.

Ejemplo 1.

Veamos un ejemplo donde se analiza una determinada variable Y para la población entre 10 y 22 años.

La recta de regresión obtenida por el método de mínimos cuadrados es:

$$y = 2,2237x + 6,449 \tag{19.1}$$

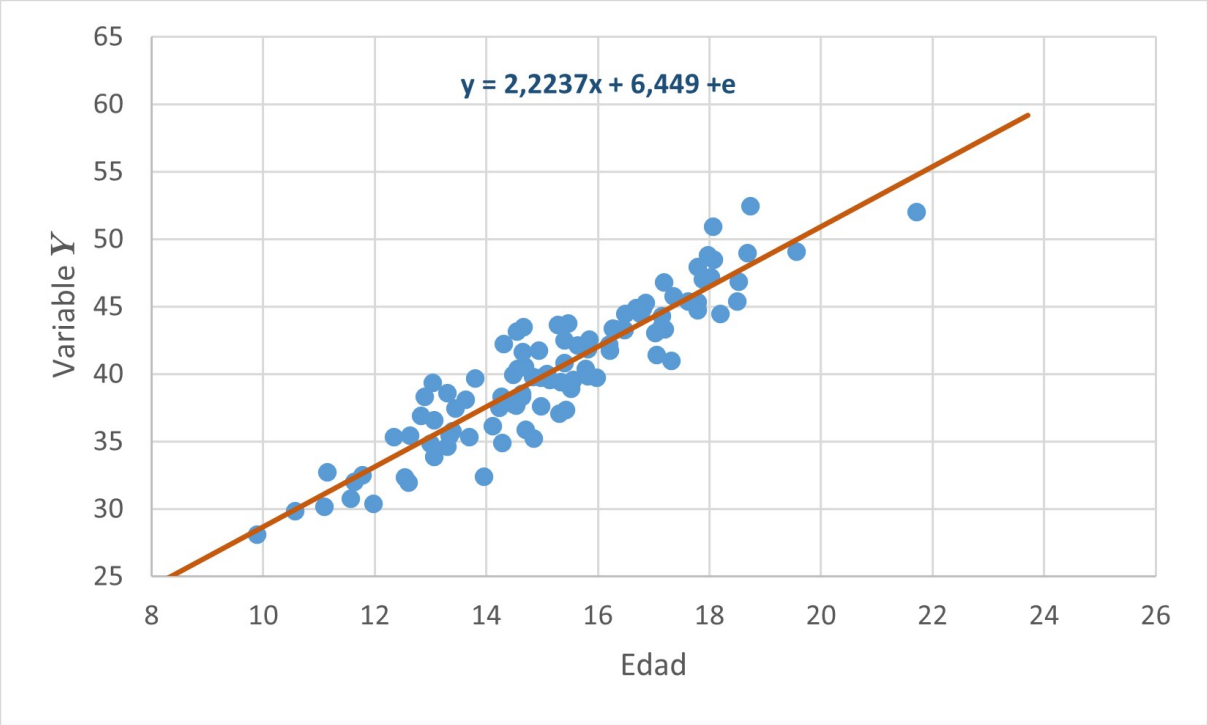


Figura 19.2

Variable truncada: la variable dependiente está truncada cuando faltan datos de la muestra (solo se han vacunado a mayores de 12 años). En este caso solo se puede modelizar para este sector de la población.

Siguiendo con el Ejemplo 1, si ahora solamente tenemos datos para los mayores de 12 años y aplicamos la técnica de mínimos cuadrados para calcular la recta de regresión, se obtiene la siguiente recta:

$$y = 2,1697x + 7,3376 + e \tag{19.2}$$

Si comparamos la ecuación de la Figura 19.2 y la de la Figura 19.3 vemos que la recta de regresión ha cambiado, esto es debido a que esta recta de regresión no contempla a los menores de 12 años, por lo tanto no se puede extrapolar la ecuación 19.3 a toda la población.

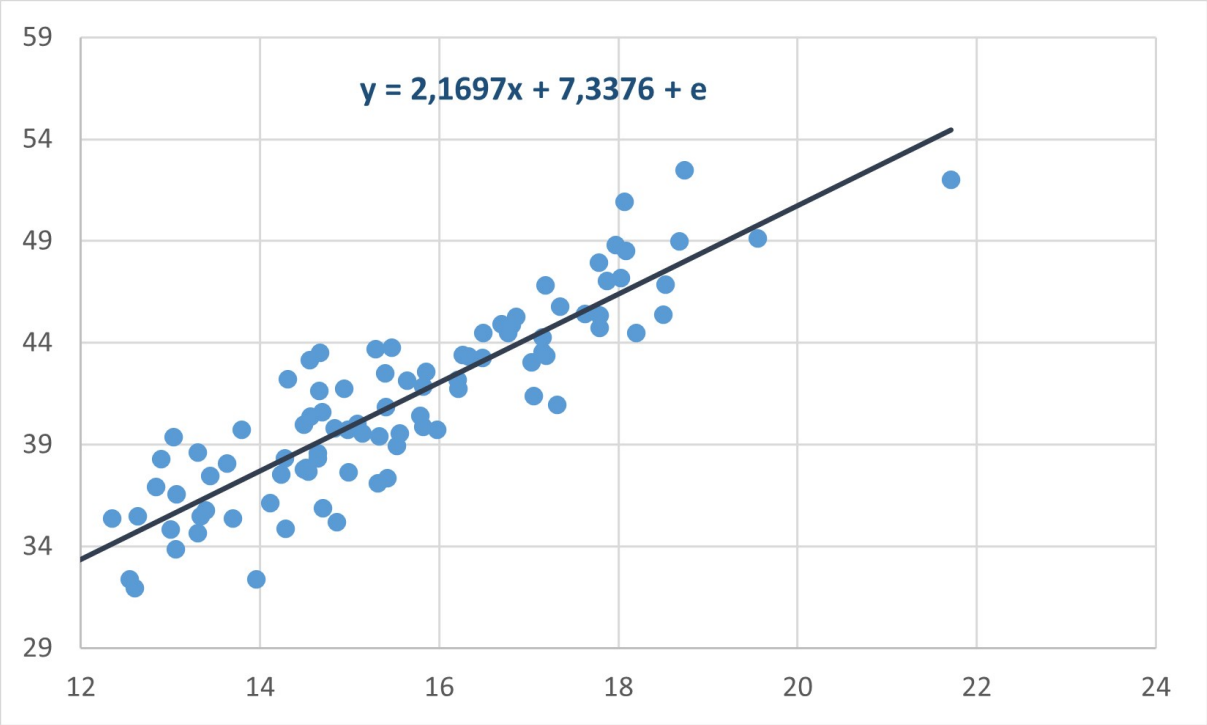


Figura 19.3

Variable dependiente censurada: a veces algunas mediciones son incompletas y cuando su valor es mayor (menor) de un cierto valor se le asigna un número simbólico.

Un ejemplo de este tipo de datos es cuando la medición se hace con un instrumento que no detecta valores mayores que un límite superior o menores que un límite inferior (termómetros, balanzas, ...) o cuando trabajamos con variables escala donde una de las clases es menor o igual a K_1 o mayor o igual a K_2.

Veamos ahora que pasa con la recta de regresión del Ejemplo 1 si la variable Y no pudiese medir los valores menores de 35. En este caso a todos los individuos cuyo valor de Y es menor que 35 se le asignará el valor 35 o el valor 0.

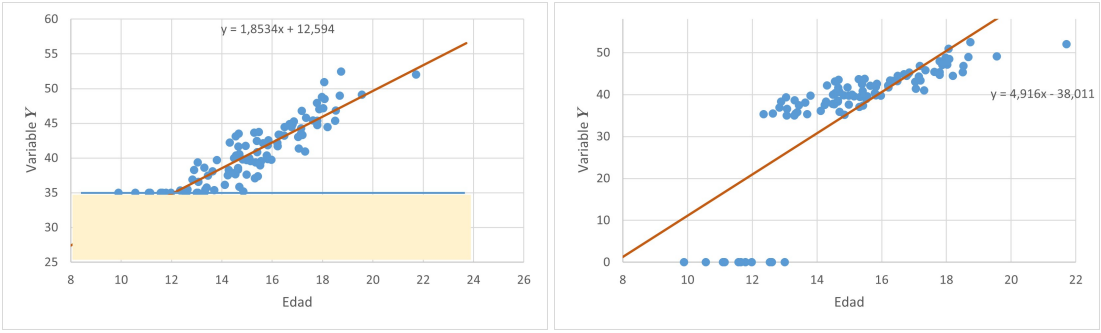
Si le asignamos el valor 35, utilizando el método de mínimos cuadrados se obtiene la siguiente ecuación:

$$y = 1,8534x + 12,594 \tag{19.3}$$

y si le asignamos el valor 0 se obtiene:

$$y = 4,916x - 38,011 \tag{19.4}$$

De nuevo si comparamos las ecuaciones de las Figuras 19.4a y 19.4b con la de la Figura 19.2 observamos que las rectas de regresión han sido modificadas, la hipótesis de que a una variable que no se puede medir exactamente se le asigne un valor mínimo o máximo distorsiona la verdadera nube de puntos.



(a) Se han sustituido los valores menores de 35 por 35

(b) Se han sustituido los valores menores de 35 por 0

Figura 19.4

Modelos Tobit (1958): surge como evolución del modelo censurado, considera que existe una variable latente Y^* no observable y, una variable Y observable formada por la parte no censurada de Y^*. El objetivo es ser capaz de estimar parámetros de Y^* empleando solo la muestra de la parte observable.

Como conclusión se puede afirmar que solamente si la variable dependiente es **Libre**, es decir, continua que puede tomar cualquier valor de $\mathbb{R}$ la técnica de mínimos cuadrados es óptima.

19.2.3 Interpretación de los coeficientes de la recta de regresión

Si se calcula la recta de regresión utilizando la técnica de mínimos cuadrados se obtiene la siguiente ecuación:

$$\hat{Y} = \bar{y} - \frac{S_{XY}}{S_X^2}\bar{x} + \frac{S_{XY}}{S_X^2}X \tag{19.5}$$

Comparando la ecuación 19.5 con la ecuación de la recta:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

se obtiene que:

$$\beta_0 = \bar{y} - \frac{S_{XY}}{S_X^2}\bar{x}$$

y

$$\beta_1 = \frac{S_{XY}}{S_X^2}$$

β_0 o termino independiente es el valor que toma la variable Y cuando la $X = 0$

y

β_1 o **coeficiente de regresión** nos va a indicar el tipo de dependencia que hay entre las dos variables, de manera que si:

- $\beta_1 > 0$ indica que la dependencia es positiva, por lo tanto si un individuo tiene mayor valor en la variable X que otro, en media, el valor de la variable Y será mayor.

  Ejemplo 2.

  Para poder estimar el peso de nuestros futuros estudiantes, hemos realizado una encuesta entre los actuales, donde se les preguntaba a cada uno su peso y altura.

  Después de comprobar que el diagrama de dispersión nos indicaba que podía haber una relación lineal entre las dos variables, calculamos la recta de regresión, obteniendo la siguiente ecuación:

  $$\hat{y_i} = 0,87x_i - 79,32$$

  Y está expresada en kg. y X en cm.

  Aunque la recta de regresión está definida para todo $\mathbb{R}$, en este caso solo tiene sentido para los valores de X que hacen que la Y sea positiva. Estudiantes que miden al menos 92cm.

  β_0: en este ejemplo no tiene sentido, pues no puede ocurrir que una persona mida cero cm.
  β_1: nos indica cuanto se incrementa el peso de un estudiante medio cuando la altura varía en un centímetro. Así si un estudiante crece un centímetro, en media, su peso se incrementará en 0,87 kg.

  Si queremos predecir el peso que va a tener un nuevo estudiante que mide 170 cm, sustituyendo en la recta de regresión obtenemos, redondeando, 69kg.

  $$\hat{y_i} = 0,87 * 170 - 79,32 = 69$$

- $\beta_1 < 0$ indica que la dependencia es negativa, es decir, que si un individuo tiene mayor valor en la variable X que otro, en media, el valor de la variable Y será menor.

  Ejemplo 3.

  Veamos ahora como predecir la temperatura media en un determinado lugar geográfico dependiendo de su latitud (el estudio se realiza el mismo día para todas las observaciones).

  Realizamos el diagrama de dispersión y a la vista del gráfico, calculamos la recta de regresión, obteniendo la siguiente ecuación:

  $$\hat{y_i} = -1,83x_i + 116,75$$

  Y está expresada en °F y X en grados, en este ejemplo solo se han tomado temperaturas en el hemisferio norte.

  β_0: en este ejemplo si se puede interpretar, pues $x = 0$ indica que estamos en el Ecuador y la temperatura media es de $116,75°F = 47°C$

β_1: nos indica cuanto disminuye, en media, la temperatura cuando ascendemos un grado en la latitud. El signo de β_1 indica el tipo de dependencia, en este caso la dependencia es inversa, a medida que subimos al Polo Norte la temperatura va disminuyendo.

Para predecir la temperatura media en la península española (latitud $27°$) , sustituimos en la recta de regresión y obtenemos una media de .

$$\hat{y_i} = -1,83 * 27 + 116,75 = 67,34°F = 19,6°C$$

- $\beta_1 = 0$ indica que la recta de regresión es: $\hat{y_i} = \beta_0$ por lo tanto la media de la variable Y permanece constante para todas las variables $Y/X = x_i$. La recta de regresión es paralela al eje de abscisas.

19.3 Coeficiente de correlación lineal y cálculo del mismo

Como ya se vió en el Tema 17, el coeficiente de correlación lineal o **coeficiente de correlación de Pearson** expresa el grado de dependecia entre las dos variables y viene dado por la expresión:

$$\rho = r = \frac{S_{XY}}{S_X S_Y}$$

Siendo: S_{XY} la covarianza entre las variables X e Y
S_X la desviación típica de la variable X
S_Y la desviación típica de la variable Y

Si conocemos la recta de regresión de Y sobre X se puede obtener el coeficiente de correlación lineal a partir de β_1.

$$\rho = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y}\frac{S_X}{S_X} = \frac{S_{XY}}{S_X^2}\frac{S_X}{S_Y} = \beta_1 \frac{S_X}{S_Y}$$

Propiedades del coeficiente de correlación

- los valores del coeficiente de correlación lineal no tiene dimensiones y van entre -1 y 1. Indicando 1 una relación funcional directa y -1 una relación funcional inversa.
- Si las variables son independientes entonces $\rho = 0$, al contrario no tiene por qué verificarse. Si $\rho = 0$ puede ocurrir que exista dependencia funcional entre las dos variables, siendo la función que relaciona a las variables no lineal.
- Si $\rho > 0$ por la relación directa que existe entre ρ y β_1, existe una dependencia positiva entre las dos variables.
- Si $\rho < 0$ existe una dependencia inversa entre X e Y, es decir, si se aumenta el valor de X también aumenta Y.

Ejemplo 4.

En el Ejemplo 2 hemos analizado la variable $Peso$ a partir de la variable $Estatura$ obteniendo la siguiente ecuación:

$$\hat{y}_i = 0,87x_i - 79,32.$$

En el Ejemplo 3 se predecía la temperatura media en un determinado lugar geográfico dependiendo de su latitud, se obtenía la siguiente ecuación:

$$\hat{y}_i = -1,83x_i + 116,75.$$

Aunque el coeficiente de correlación lineal solamente nos indica el grado de dependencia lineal entre dos variables sin distinguir entre variable dependiente y variable independiente, también se puede utilizar para saber cual de los dos modelos realiza mejores predicciones.

El coeficiente de correlación en el caso de Peso-Altura es $0,76$, al ser positivo nos indica que la dependencia es directa. A mayor altura, mayor peso.

En el caso Temperatura-Latitud el coeficiente de correlación lineal es $-0,90$, en este caso al ser negativo nos indica una dependencia inversa. Mayor latitud, en media, implica menor temperatura.

Si ahora comparamos en valor absoluto los dos coeficientes, la dependencia lineal es más fuerte entre el par de variables Temperatura-Latitud que entre Peso-Altura. Esto no implica que el primer par de variables estén más relacionadas que el segundo, pues como ya se ha señalado anteriormente, la falta de dependencia lineal no implica que exista independencia entre las variables.

Sin embargo, si utilizamos la recta de regresión para hacer predicciones de las variables dependientes, lo que si se puede afirmar es que, las predicciones para la variable Temperatura serán más aproximadas que para la variable Peso.

Si escribimos:

$$r^2 = 1 - \frac{S_e^2}{S_Y^2}$$

esta expresión nos indica que el coeficiente de correlación lineal al cuadrado es el porcentaje de varianza de la variable dependiente que está explica por la variable dependiente.

	r	r^2	Reducción de la Varianza
Peso-Altura	0,76	0,58	$58\,\%$
Temperatura-Latitud	0,90	0,81	$81\,\%$

Tabla 19.1

En nuestro caso, la varianza de la variable Peso se reduce en un 58 % y el de la Temperatura en un 81 %.

- Cuando en una muestra hay subgrupos (sexo, provincias, ...) podemos encontrar que el coeficiente de correlación sea distinto en cada subgrupo y muy diferente del calculdo para todos los individuos.

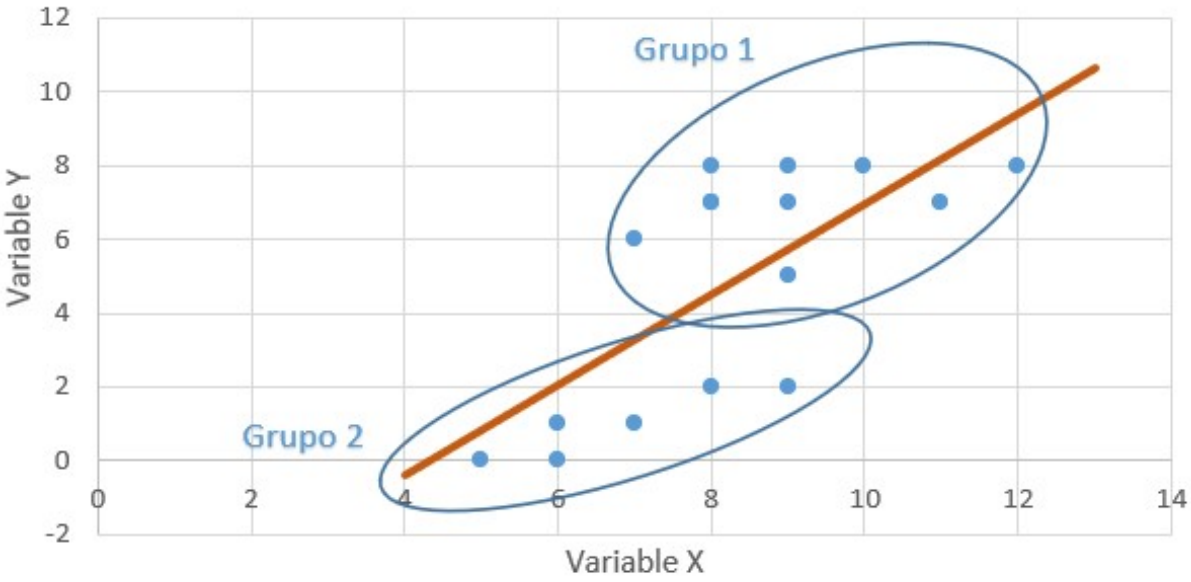


Figura 19.5

x_i	y_i	$x_i y_i$	y_i^2	x_i^2
12	8	96	64	144
10	8	80	64	100
11	7	77	49	121
8	7	56	49	64
7	6	42	36	49
9	8	72	64	81
8	7	56	49	64
9	7	63	49	81
9	5	45	25	81
8	8	64	64	64
Suma				
9	7	65	51	85

Tabla 19.2: Datos par el Grupo 1

Para el Grupo 1 se obtiene: $S_{XY} = 0,49$, $S_Y = 0,94$, $S_X = 1,45$
Por lo tanto el coeficiente de correlación lineal es $r = \frac{S_{XY}}{S_X S_Y} = 0,36$.

Para el Grupo 2

x_i	y_i	x_iy_i	y_i^2	x_i^2
6	0	0	0	36
5	0	0	0	25
7	1	7	1	49
6	1	6	1	36
9	2	18	4	81
8	2	16	4	64
Suma				
9	7	65	51	85

Tabla 19.3: Datos par el Grupo 2

$S_{XY} = 1$, $S_Y = 0,81$, $S_X = 1,34$ y por tanto, el coeficiente de correlación lineal es $r = 0,91$.

Como podemos observar en el Grupo 1 existe menor dependencia lineal que en el Grupo 2.

Veamos ahora que ocurre si calculamos el coeficiente de correlación lineal para todos los datos.

En este caso $S_{XY} = 3,92$, $S_Y = 3,09$, $S_X = 1,79$ y el coeficiente de correlación lineal $r = 0,71$.

En el caso que el diagrama de dispersión nos indique que la muestra puede estar formada por distintos grupos, se debe realizar el modelo de regresión para cada uno de los grupos y posteriormente evaluar si en todos los grupos la recta de regresión puede considerarse la misma.

- Valores atípicos, el coeficiente de correlación es muy sensible a los valores extremos debido a que para su cálculo se utiliza la media aritmética.

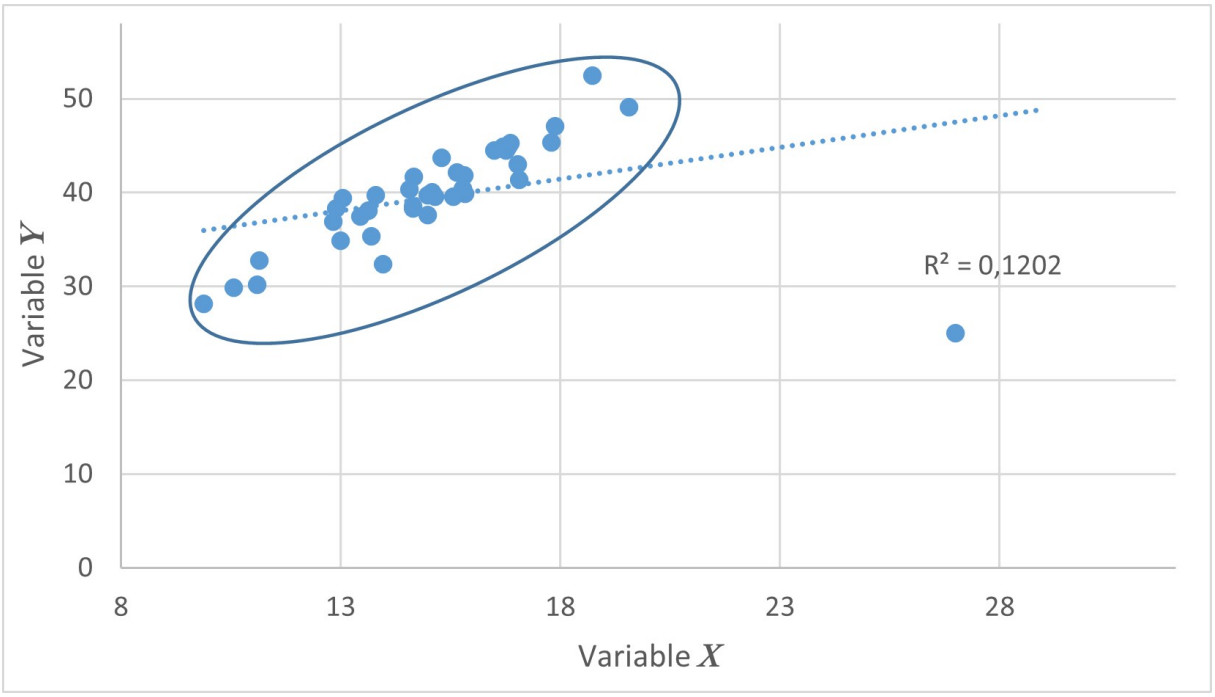


Figura 19.6

En la Figura 19.6 se puede observar como el individuo que no pertenece al grupo modifica sustancialmente la recta de regresión y el coeficiente de correlación.

19.4 Posiciones de las rectas de regresión según el valor del coeficiente de correlación

La recta de regresión lineal de Y sobre X, donde Y es la variable dependiente y X la variable independiente, hemos visto que viene expresada como:

$$\hat{Y} - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

Para obtener la recta de regresión X sobre Y se procede de forma análoga, repitiendo el método de Mínimos Cuadrados sobre la media de los cuadrados de los residuos, que en este caso, los residuos de la variable X.

$$e_i^2 = (x_i - \hat{x}_i)^2$$

Minimizamos la suma de los residuos y obtenemos la ecuación de la recta de regresión:

$$\hat{X} - \bar{x} = \frac{S_{XY}}{S_Y^2}(Y - \bar{y})$$

Solamente en el caso de que la dependencia lineal sea exacta ($r = \pm 1$) las dos rectas de regresión coinciden.

Ejemplo 5.

A partir de los datos de la siguiente tabla:

X	9	10	11	12	13	14	15
Y	29	31	32	37	36	40	43

Tabla 19.4: Cálculos intermedios

se van a calcular las dos rectas de regresión.

Para ello construimos la siguiente tabla:

X	Y	X^2	Y^2	$X * Y$
9	29	81	841	261
10	31	100	961	310
11	32	121	1024	352
12	37	144	1369	444
13	36	169	1296	468
14	40	196	1600	560
15	43	225	1849	645
Suma				
84	248	1036	8940	3040

Tabla 19.5: Cálculos intermedios

y a partir de ella se calculan las propiedades de las variables X e Y que se necesitan para obtener las rectas de regresión.

Vector de medias:

$$(\bar{x}, \bar{y}) = (\frac{84}{7}, \frac{248}{7}) = (12; 35, 43)$$

Covarianza de X e Y:

$$S_{XY} = \frac{3040}{7} - 12 * 35, 43 = 9, 14$$

Varianza de X:

$$S_X^2 = \frac{1036}{7} - 12^2 = 4$$

Varianza de Y:

$$S_Y^2 = \frac{8940}{7} - 35, 43^2 = 21, 96$$

Coeficiente de regresión para la recta de Y sobre X:

$$\beta_1 = \frac{9,14}{4} = 2,29$$

Coeficiente de regresión para la recta de X sobre Y:

$$\beta_1' = \frac{9,14}{21,96} = 0,42$$

Sabiendo que la recta de regresión de Y sobre X es:

$$\hat{Y} - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

La ecuación de la recta de regresión de Y sobre X es:

$$\hat{Y} - 35,43 = 2,29(X - 12)$$

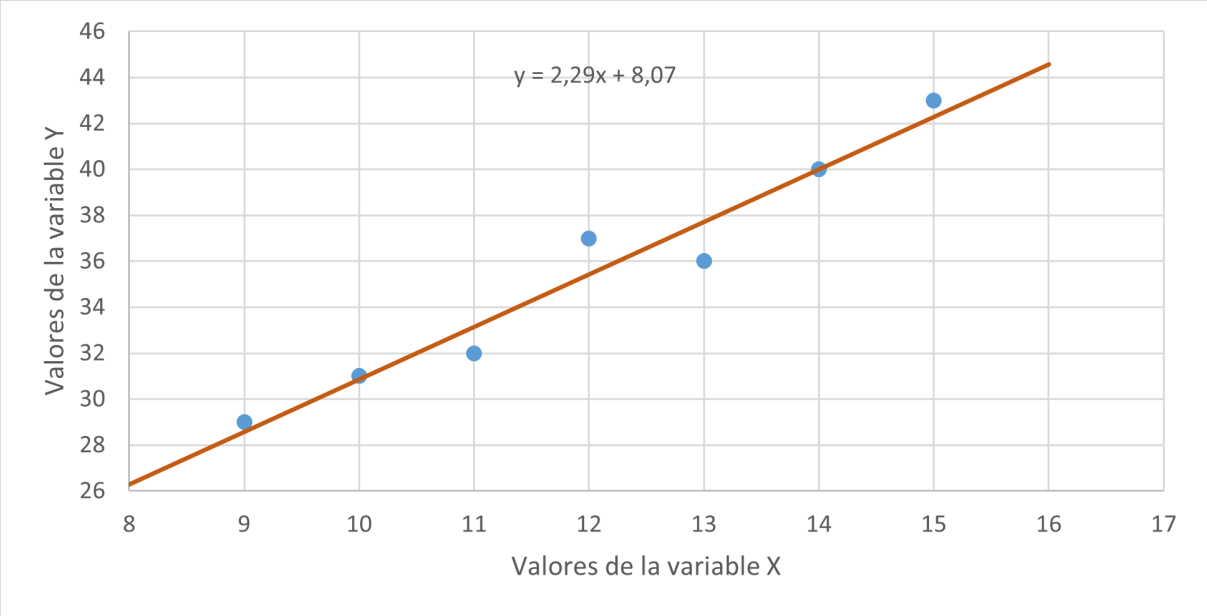


Figura 19.7

La recta de regresión de X sobre Y viene dada por:

$$\hat{X} - \bar{x} = \frac{S_{XY}}{S_Y^2}(Y - \bar{y}) = \beta_1'(Y - \bar{y})$$

Sustituyendo los parámetros de la ecuación, se obtiene la siguiente recta:

$$\hat{X} - 12 = 0,42(Y - 35,43)$$

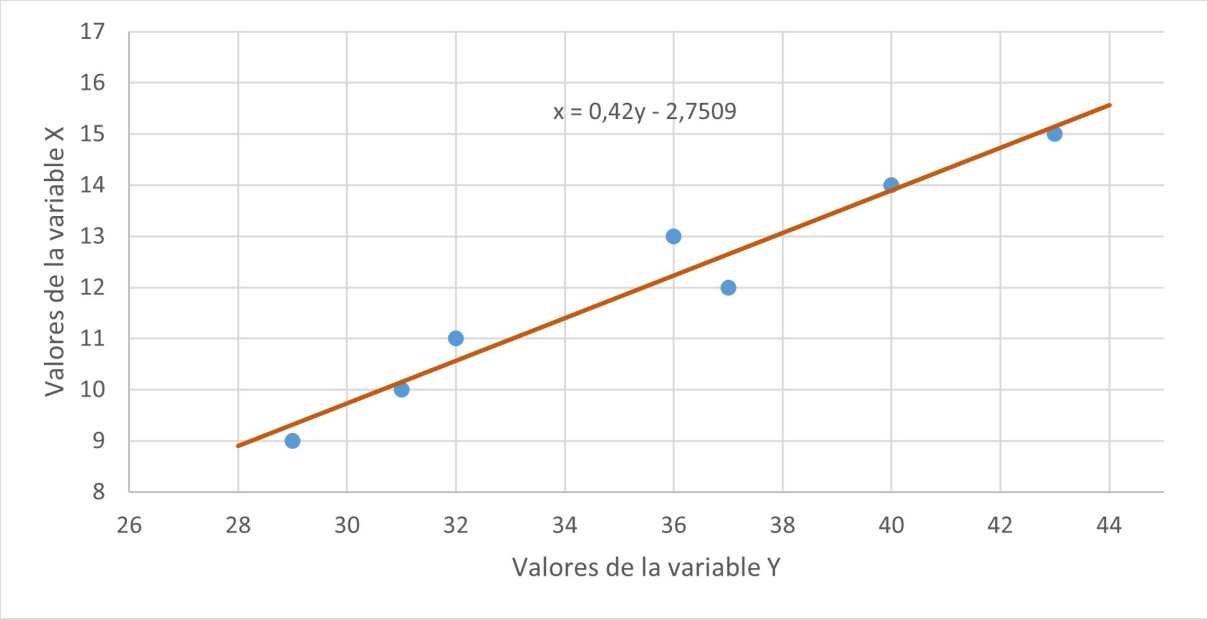


Figura 19.8

Si se dibujan las dos rectas de regresión sobre el mismo gráfico se obtiene la siguiente representación:

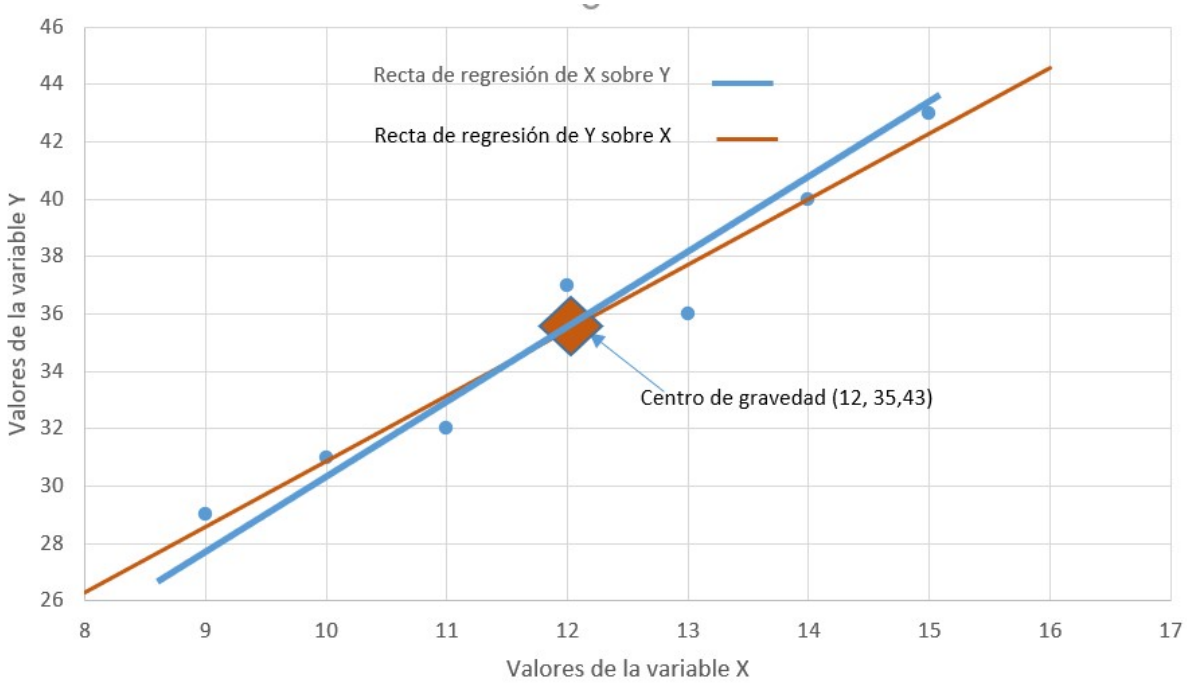


Figura 19.9

Como podemos observar en la Figura 19.9 las dos rectas de regresión se cortan en el centro de gravedad $(\bar{x}, \bar{y}) = (12; 35, 43)$.

Para que las dos rectas fuesen coincidentes las dos pendientes deberían de ser inversas.

Recta de regresión de Y sobre X

$$\hat{Y} - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x}).$$

Recta de regresión de X sobre Y

$$\hat{X} - \bar{x} = \frac{S_{XY}}{S_Y^2}(Y - \bar{y})$$

despejando la variable Y para ponerlas las dos de la misma forma, obtenemos:

$$\hat{Y} - \bar{y} = \frac{S_Y^2}{S_{XY}}(X - \bar{x})$$

Por lo tanto para que sean coincidentes

$$\frac{S_Y^2}{S_{XY}} = \frac{S_{XY}}{S_X^2}$$

Esta igualdad solamente se verifica si $r^2 = 1$, es decir, si existe dependencia funcional lineal.

En nuestro caso $r^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2} = 0,952$.

Propiedades de las rectas de regresión

1. Todas las rectas de regresión pasan por el punto $(\bar{x}, \bar{y})$. A este punto se le denomina **centro de gravedad**. La mejor estimación para la variable Y cuando $X = \bar{x}$ es su media ($\bar{y}$) y viceversa, la mejor estimación para la variable X cuando $Y = \bar{y}$ es $\bar{x}$.
2. Si la correlación lineal es cero, las rectas de regresión son perpendiculares.

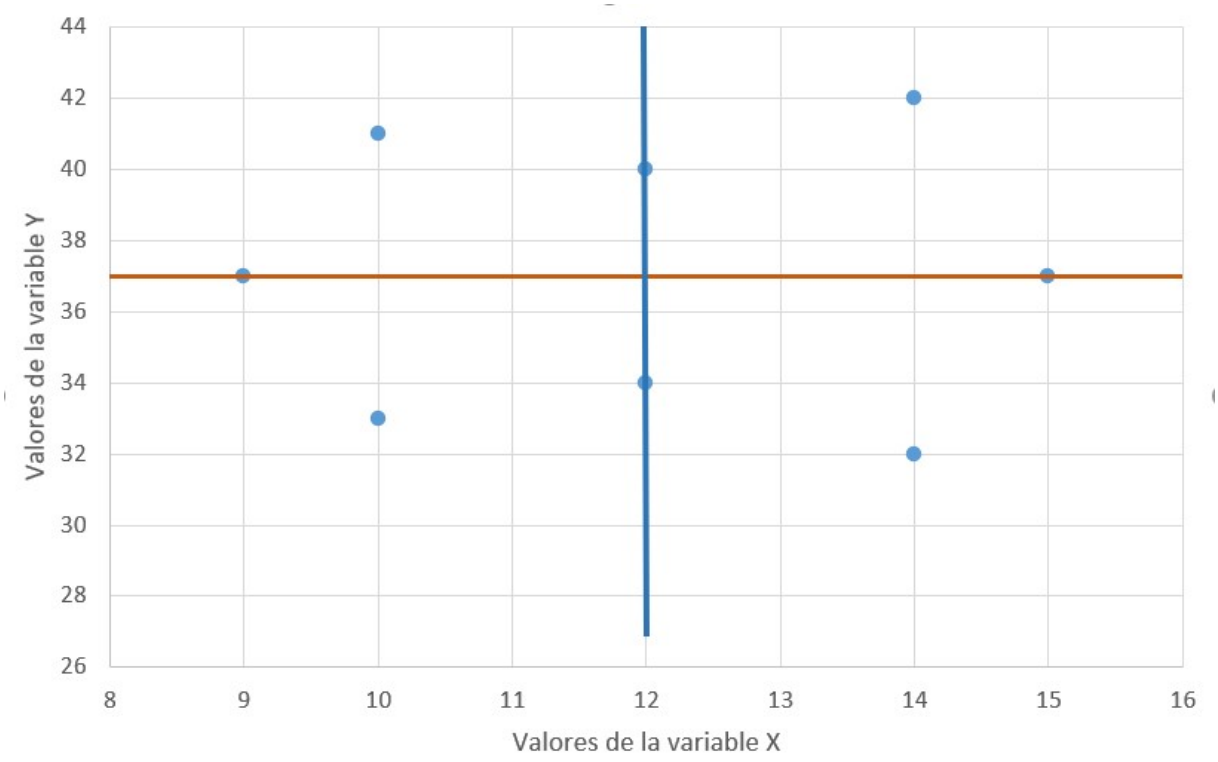


Figura 19.10

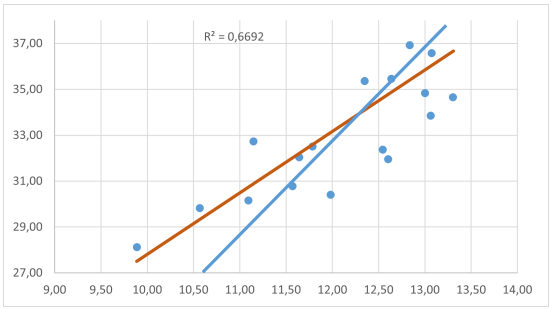
Que la correlación lineal sea cero indica que son variables linealmente independientes y por lo tanto la mejor estimación para la variable Y será su media ($\bar{y}$), independientemente del valor que tome la variable X. la recta de regresión de Y sobre X es paralela al eje de abscisas, $\hat{Y} = \bar{y}$.
Si la variable que queremos estimar es la variable X, calculamos la recta de regresión de X sobre Y y como la $S_{XY} = 0$ volvemos a obtener una recta paralela al eje, en este caso paralela al eje de ordenadas, $\hat{X} = \bar{x}$. La mejor estimación para la variable X, en este caso, es su media $\bar{x}$ para cualquier valor de Y.

3. Si $r = \pm 1$ las dos rectas de regresión son coincidentes.
   En el caso de que $r = 1$ la dependencia es positiva y las rectas de regresión son crecientes.
   Si $r = -1$ la dependencia es negativa y las rectas de regresión son decrecientes.

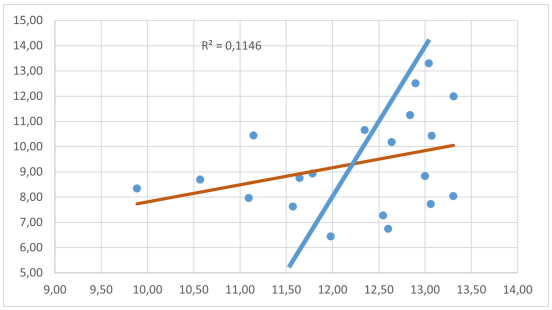
4. Si $|r| \neq (0, 1)$, las dos rectas de regresión son no coincidentes. Cuanto mayor sea el coeficiente de correlación menor será el ángulo entre las dos rectas de regresión como se puede observar en la Figura 19.11.

5. El producto de los dos coeficientes de correlación es igual al coeficiente de determinación. Coeficiente de regresión para la recta de Y sobre X:

$$\beta_1 = \frac{S_{XY}}{S_X^2}$$



(a) Rectas de regresión con r próximo a 1



(b) Rectas de regresión con r próximo a 0

Figura 19.11

Coeficiente de regresión para la recta de X sobre Y:

$$\beta_1' = \frac{S_{XY}}{S_Y^2}$$

Si multiplicamos $\beta_1 * \beta_1'$ se obtiene:

$$\beta_1 * \beta_1' = \frac{S_{XY}^2}{S_X^2 * S_Y^2} = R^2$$

6. Se cumple las siguientes relaciones:

$$\beta_1 = r\frac{S_X}{S_Y}$$

y

$$\beta_1' = r\frac{S_Y}{S_X}$$

Ejemplo 6.

Analicemos los datos que corresponden a un experimento donde se mide la longitud de distintos muelles dependiendo del peso que se ponga en el extremo.

Peso	0	2	5	11	17	2	5	11	15	17	16
Longitud Muelle	8	28	31	49	72	19	21	25	62	35	30

Tabla 19.6: Datos del experimento

Lo primero representamos los datos del experimento

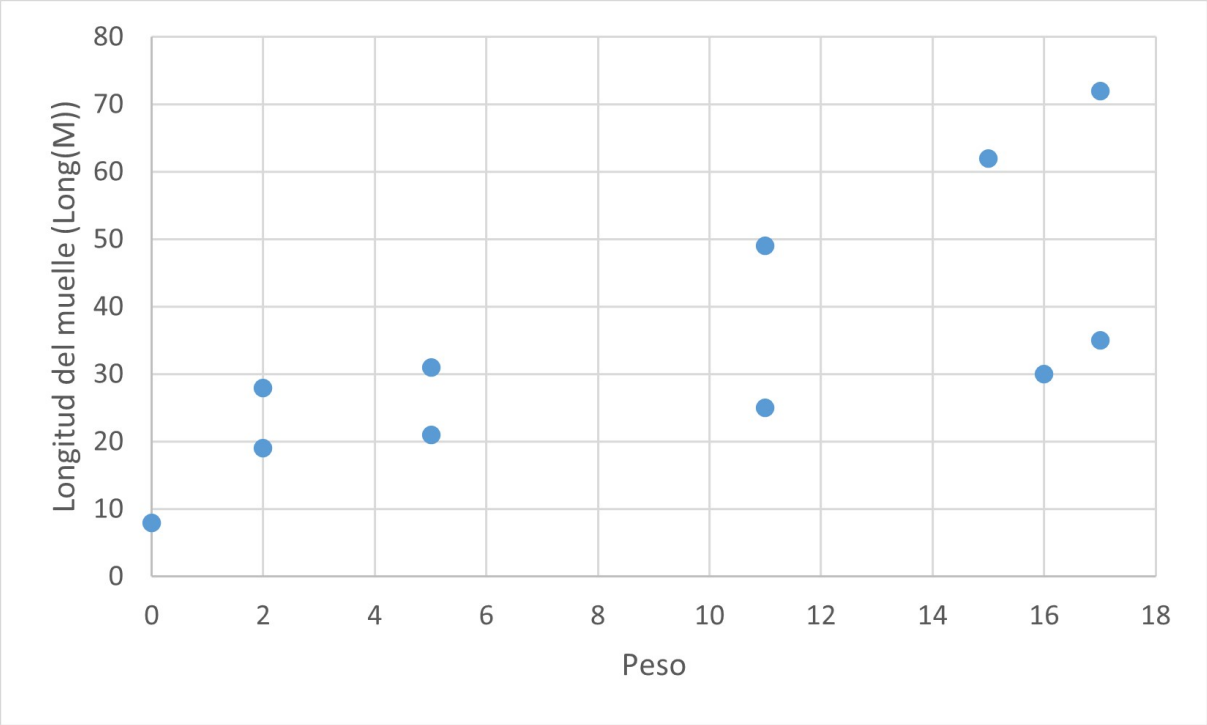


Figura 19.12: Diagrama de dispersión

En la Figura 19.12 se puede observar una dependencia directa, para asegurarnos que existe una dependencia lineal calculamos el coeficiente de correlación lineal.

Peso	Long(M)	Peso*Long(M)	$Peso^2$	$Long(M)^2$
0	8	0	0	64
2	28	56	4	784
5	31	155	25	961
11	49	539	121	2401
17	72	1224	289	5184
2	19	38	4	361
5	21	105	25	441
11	25	275	121	625
15	62	930	225	3844
17	35	595	289	1225
16	30	480	256	900
Medias				
9,18	34,55	399,73	123,55	1526,36

Tabla 19.7

A partir de los datos de la Tabla 19.7 calculamos:
$S(Peso * Long(M)) = 82,54$, $S(Long(M)) = 18,25$, $S(Peso) = 6,26$ y $r = 0,72$

Lo que nos indica que la dependencia lineal entre las variables es buena.

Calculamos el Coeficiente de Determinación: $R^2 = 0,52$ por lo tanto, más del 50 % de la varianza de una de las variables es explicada por la otra a través de un modelos lineal.

A continuación calculemos las dos rectas de regresión.

Si denominamos Y a la variable $Peso$ y X a la variable Longitud del Muelle ($Long(M)$).

La recta de regresión del $Peso$ sobre $Long(M)$ viene dada por la expresión:

$$\hat{Y} - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x}).$$

Si sustituimos, obtenemos la siguiente ecuación de la recta

$$\hat{Y} - 9,18 = \frac{82,54}{332,98}(X - 34,55).$$

La otra recta de regresión de $Long(M)$ sobre $Peso$ es:

$$\hat{X} - \bar{x} = \frac{S_{XY}}{S_Y^2}(Y - \bar{y})$$

$$\hat{X} - 34,55 = \frac{82,54}{39,24}(Y - 9,18)$$

Como se puede observar las dos rectas de regresión no son coincidentes, esto es debido a que la relación de dependencia entre ellas no es perfecta, $r \neq 0$

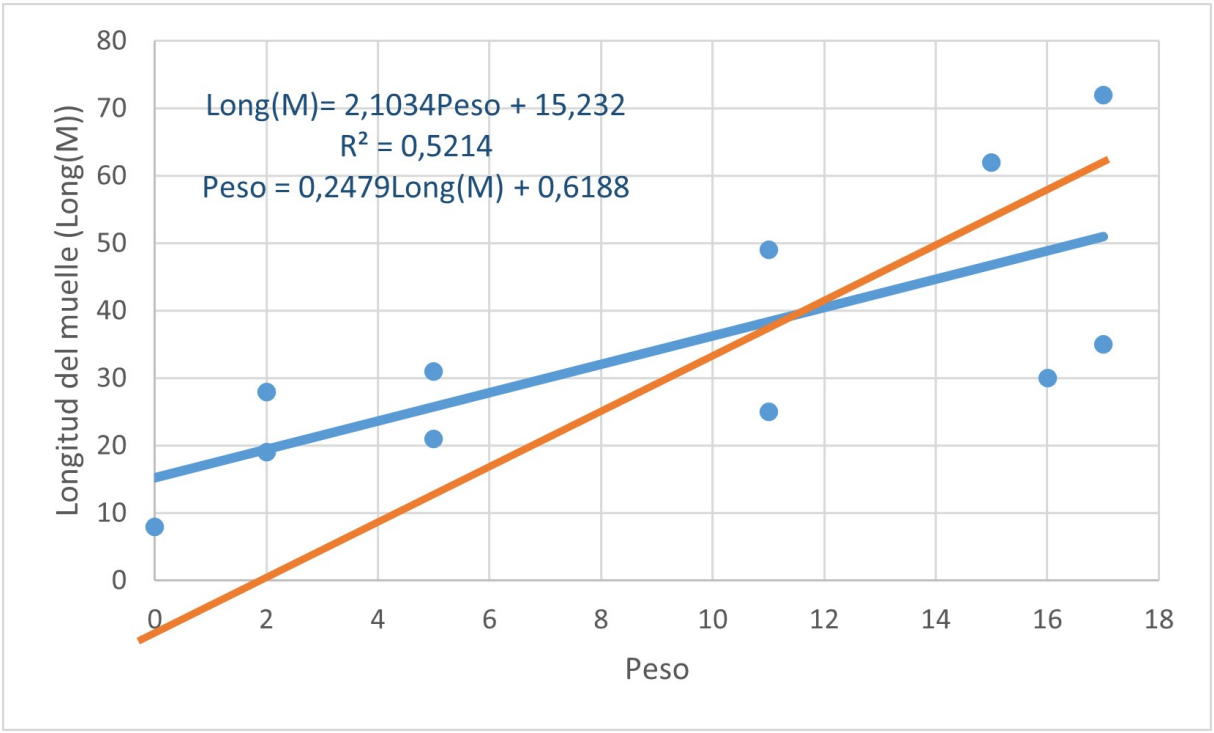


Figura 19.13: Diagrama de dispersión

Bibliografía

Montiel Torres, AM, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson (página 65).

Peña, D (2002). *Regresión y Análisis de Experimentos*. Alianza Editorial (página 65).

Camacho, Carlos, AM López y MA Arias (2006). "Regresión lineal simple". En: *de Apuntes no publicados de la asignatura Análisis de datos II de la licenciatura de Psicología, Universidad de Sevilla* (página 65).

Montgomery, Douglas, Elizabeth Peck y Geoffrey Vining (2006). *Introducción al análisis de regresión lineal*. México: Limusa Wiley (página 65).

Tema 20

Distribución de frecuencias n-dimensional. Regresión múltiple. Regresión lineal múltiple, correlación lineal múltiple y parcial. Multicolinealidad.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

Roberto Montero Granados (2016). "Modelos de regresión lineal múltiple". En: *Granada, España: Departamento de Economía Aplicada, Universidad de Granada*

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

20.1 Introducción

En este tema vamos a aprender cómo analizar la relación simultánea de más de dos variables. La técnica que emplearemos será la regresión lineal múltiple. También introduciremos los estadísticos: coeficiente de determinación múltiple, coeficiente de correlación múltiple y parcial, etc.

Trabajaremos con tablas donde tengamos múltiples variables y analizaremos la relación entre todas las variables o algunas de ellas. Por ejemplo, se puede disponer de tablas como la siguiente:

ID	X1	X2	X3	X4	X5	X6	X7	X8
1	283	299.935	4	44000	0	283.139	304.935	84881.605
2	280	742.44	3	44394	1	560.478	1451.88	207883.2
3	2	396.775	4	43539	1	6.299	1124.325	793.55
4	270	668.465	2	44146	0	1080.396	2595.86	180485.55
5	66	174.85	6	43647	1	330.379	884.25	11540.1
6	312	689.985	7	43882	0	1872.441	4048.91	215275.32
7	394	437.125	5	43888	0	2758.926	3100.875	172227.25
8	178	326.835	4	44122	0	1424.891	2530.68	58176.63
...	...	...	...	...	...	...	...	...

20.2 Distribución de frecuencias n-dimensional

En muchas ocasiones interesa estudiar el comportamiento de más de una variable aleatoria en una población. El análisis no solo debe ser unidimensional (cada variable por separado), también es necesario analizar su comportamiento conjunto con el fin de determinar la influencia de una en otra u otras.

Para realizar el estudio conjunto de las variables aleatorias, lo primero que vamos a definir es la **distribución de frecuencias multidimensional**.

$Individuo$	X_1	X_2	X_3	...	X_k
1	x_{11}	x_{21}	x_{31}		x_{k1}
2	x_{12}	x_{22}	x_{32}	...	x_{k2}
3	x_{13}	x_{23}	x_{33}	...	x_{k3}
...	...	...	...	...	...
i	x_{1i}	x_{2i}	x_{3i}	...	x_{ki}
...	...	...	...	...	...
N	x_{1N}	x_{2N}	x_{3N}	...	x_{kN}

Tabla 20.1: Valores de las variables X_1, ..., X_k

Estas variables pueden ser cualitativas o cuantitativas, en este tema nos vamos a centrar en las cuantitativas puesto que para realizar una regresión se necesitan características como la media, la varianza o la correlación lineal. Estas características se pueden calcular tanto en el caso de variables aleatorias continuas como en el de variables aleatorias discretas.

Ejemplo 1. Un grupo de cuatro estudiantes han aportado la siguiente información para un estudio:

Horas de estudio	Consumo de cafeína	Peso
1	0,25	58
3	0,92	76
5	1,43	61
8	2,15	55

La tabla anterior consta de 3 variables y 4 observaciones. Por tanto, la distribución es tridimensional.

Representación gráfica de distribuciones n-dimensionales

La representación gráfica de distribiciones n-dimensionales supone un aumento en la complejidad en cuanto al número de ejes necesarios para representar todas las variables. Por este motivo, la representación se suele hacer a nivel bidimensional, eligiendo dos variables del conjunto de variables; o a nivel tridimensional, eligiendo tres variables.

Por ejemplo, la representación de los datos del Ejemplo 1 sería la siguiente:

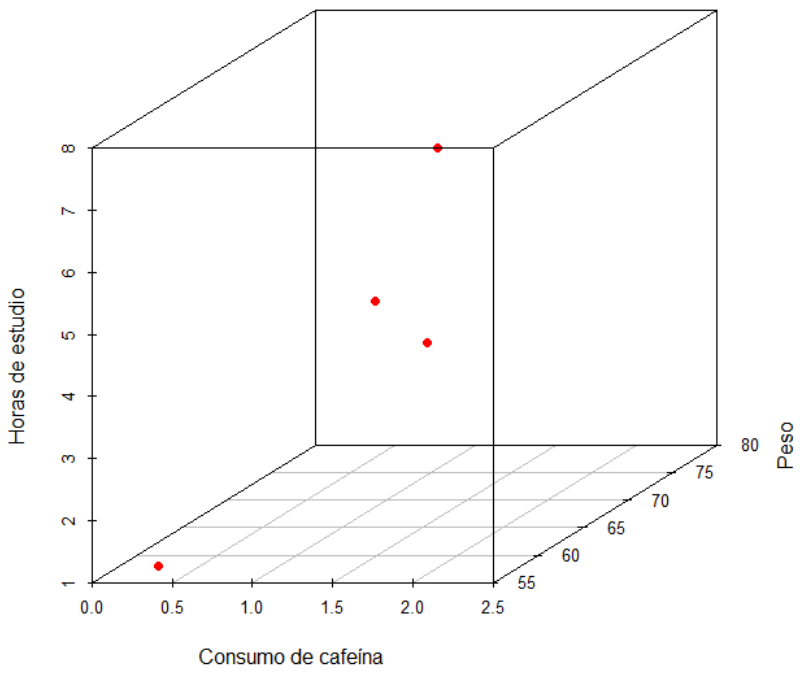


Figura 20.1: Representación de las horas de estudio (X_1), consumo de cafeína (X_2) y peso (X_3)

Alternativamente, se puede fijar una tercera variable (o varias variables) y para ese valor concreto de la tercera variable (o los valores fijados para varias variables) se construyen gráficos bidimensionales con otras variables. Y lo mismo para los gráficos trimidimensionales.

Ejemplo 2. Se han medido cinco variables para medir el funcionamiento de las máquinas de una fábrica. Los datos aparecen en la siguiente tabla:

X_1	X_2	X_3	X_4	X_5
15	10,25	158	0,345	23
23	0,92	476	0,275	62
65	1,43	561	0,149	29
48	2,15	755	0,901	70
15	10,25	158	0,345	23
15	8,25	145	0,290	30
23	0,92	476	0,275	62
65	1,43	561	0,149	29
48	2,15	755	0,901	70
15	10,25	158	0,345	23
23	0,92	476	0,275	62

La tabla anterior consta de 5 variables y 11 observaciones.

También podríamos recoger los datos del siguiente modo:

X_1	X_2	X_3	X_4	X_5	n_{ijklm}
15	10,25	158	0,345	23	3
23	0,92	476	0,275	62	3
65	1,43	561	0,149	29	2
48	2,15	755	0,901	70	2
15	8,25	145	0,290	30	1

Todo lo estudiado para las distribuciones unidimensionales y bidimensionales puede aplicarse a los datos anteriores. Es decir, podemos calcular la media de la variable X_1 y el momento ordinario bidimensional de orden uno para X_1 y orden uno para X_2.

Los estadísticos unidimensional y bidimensional serían:

$$\bar{X_1} = 32,27;$$

$$a_{11}(X_1, X_2) = 94,62.$$

La representación conjunta de las variables X_1, X_2 y X_5 sería la siguiente:

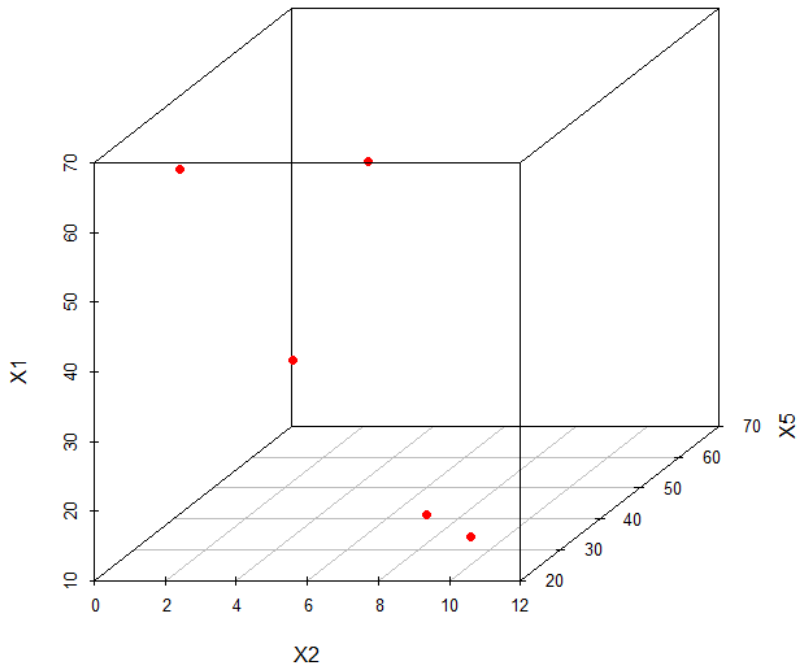


Figura 20.2: Representación de las variables X_1, X_2 y X_5

Por último, también es posible representar la relación de tres variables fijando los valores de una cuarta variable. En las siguientes figuras representamos la relación de Y, X_1 y X_2 en función de los valores de la variable X_3.

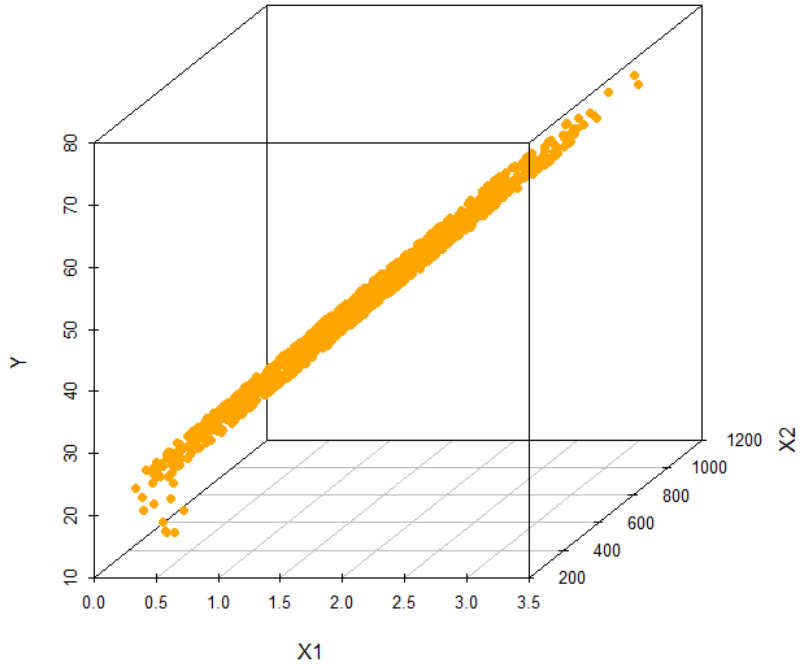


Figura 20.3: Representación de la relación de las variables Y, X_1 y X_2 cuando $X_3 = 0$

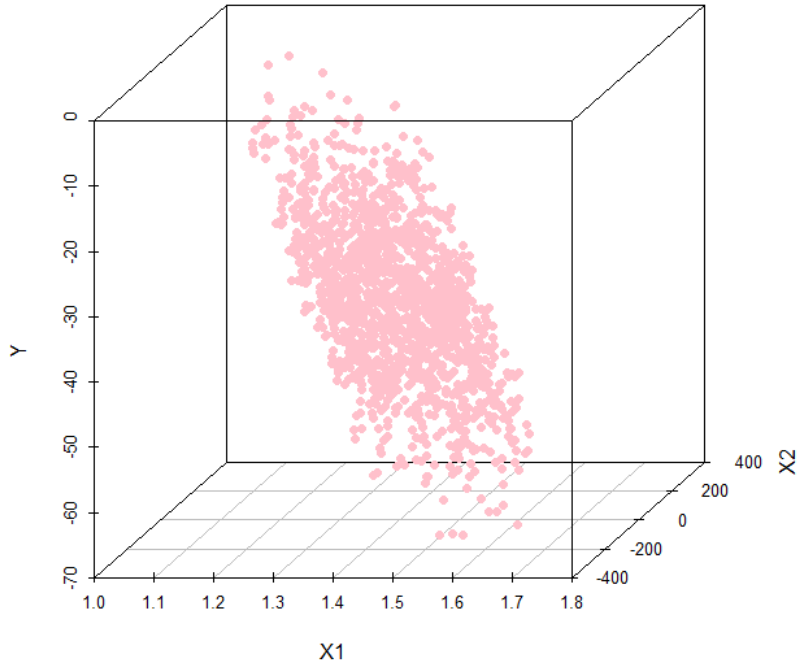


Figura 20.4: Representación de la relación de las variables Y, X_1 y X_2 cuando $X_3 > 0$

20.3 Regresión múltiple

El tipo de variable es importante sobre todo en el caso de la variable dependiente, pues dependiendo del tipo se utilizará uno u otro modelo de regresión.

Tipo de variable dependiente	Modelo
Continua	Lineal
Dicotómica	Logit o probit
Recuento	Poison o Binomial
Factor ordenado	Logit o probit ordenada
Porcentaje	Regresión fraccional

Tabla 20.2: Distintos modelos de regresión

La interpretación de los coeficientes parciales de regresión tendrán distintas interpretaciones dependiendo del tipo de variable.

Aunque existen muchas técnicas de regresión en función del tipo de variables y de la forma funcional supuesta entre ellas, las más elementales (aunque las más potentes

en el sentido de que se puede obtener más información) son las lineales. La regresión lineal supone que la relación entre dos variables tiene una forma lineal (o linealizable mediante alguna transformación de las variables).

20.3.1 Regresión lineal múltiple

La regresión lineal simple estima una variable a partir de otra, sin embargo, estos modelos suelen ser insuficiente para entender fenómenos mínimamente complejos en la que influyen más de dos variables. En el modelo de regresión lineal múltiple suponemos que más de una variable tiene influencia o está correlacionada con el valor de una tercera variable. Por ejemplo, en el peso de una persona pueden influir su edad, el género o la estatura entre otras.

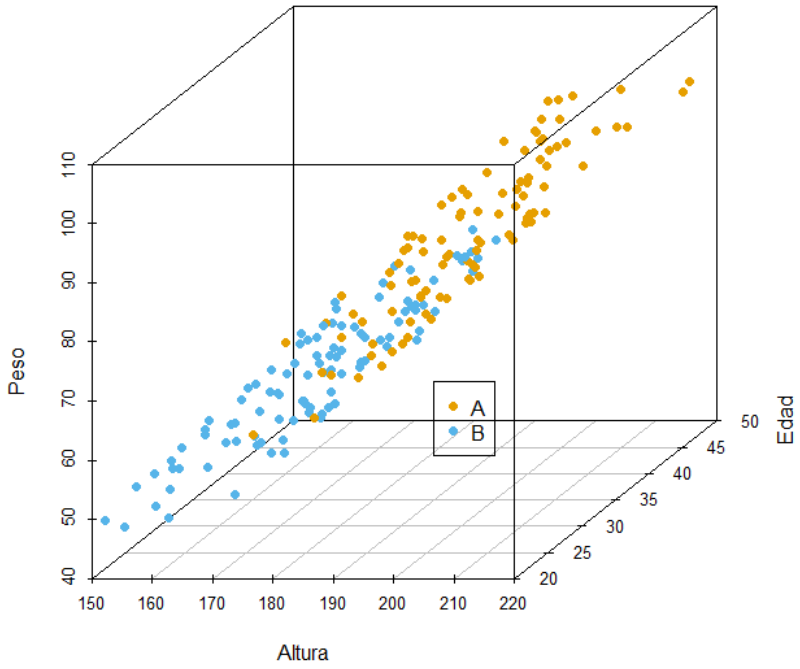


Figura 20.5: Representación del peso, altura y edad para los grupos A y B

La regresión lineal múltiple es una extensión de la regresión lineal simple. Si se utiliza la regresión para evaluar la influencia que tienen los predictores sobre ella, debemos mostrar cautela para no malinterpretar causa-efecto.

La regresión múltiple genera un modelo lineal entre una variable dependiente (variable endógena) Y y un conjunto de variables independientes (variables exógenas o predictores) $X_1, X_2, X_3, ..., X_k$ de forma que cada observación y_i se puede poner como

combinación lineal de las observaciones de los predictores más un residuo.

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$

siendo:
β_0: es la ordenada en el origen, es decir, el valor de la variable dependiente Y cuando todos los predictores son cero.

β_i: es la variación media de la variable Y cuando se incrementa en una unidad de la variable predictora X_i y se mantienen constantes el resto de variables. Se conocen como coeficiente parcial de regresión.

$\hat{y}_i$: la estimación de la variable Y para el individuo i-ésimo a través del modelo de regresión.

e_i: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo para cada individuo.

Es importante tener en cuenta que la magnitud de cada coeficiente parcial de regresión (β_i) no está asociada a la importancia del predictor X_i.

Los modelos de regresión múltiple requieren las mismas hipótesis que la regresión lineal simple y además entre los predictores debe existir independencia. Por lo tanto es importante comprobar las hipótesis de heterocedaticidad, multicolinealidad y especificidad.

20.3.2 Regresión lineal múltiple para 3 variables

Dada una distribución conjunta de Y, X_1 y X_2, se ajusta el plano

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Los parámetros que nos interesan determinar son:

$$\beta_0,$$
$$\beta_1,$$
$$\beta_2.$$

Para ello, al igual que en el caso de regresión lineal simple, nos interesa minimizar la nube puntos por mínimos cuadrados de la siguiente diferencia:

$$y_i - \hat{y}_i$$

Es decir, se busca minimizar la suma de cuadrados de la diferencia entre el valor real y el predicho:

$$\sum (y_i - \hat{y}_i)^2.$$

Y los valores β_0, β_1, y β_2 se obtienen resolviendo el siguiente sistema matricial:

$$\begin{pmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{pmatrix} = \begin{pmatrix} N & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Y si despejamos las betas y la matriz es invertible, la solución del sistema anterior es:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} N & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{pmatrix}$$

Ahora operamos:

$$\begin{aligned} \beta_1 &= \frac{S_{X_2}^2 S_{X_1 Y} - S_{X_1 X_2} S_{X_2 Y}}{S_{X_1}^2 S_{X_2}^2 - S_{X_1 X_2}^2} \\ \beta_2 &= \frac{S_{X_1}^2 S_{X_2 Y} - S_{X_1 X_2} S_{X_1 Y}}{S_{X_1}^2 S_{X_2}^2 - S_{X_1 X_2}^2} \\ \beta_0 &= \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 \end{aligned}$$

Aquí, se debe recordar la matriz de varianzas y covarianzas de las variables tiene la siguiente configuración:

$$S_{X_1 X_2 Y} = \begin{pmatrix} S_{X_1}^2 & S_{X_1 X_2} & S_{X_1 Y} \\ S_{X_2 X_1} & S_{X_2}^2 & S_{X_2 Y} \\ S_{Y X_1} & S_{Y X_2} & S_Y^2 \end{pmatrix}$$

Sustituyendo el valor de β_0 en la ecuación $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, se tiene que

$$y = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 + \beta_1 x_1 + \beta_2 x_2,$$

y operando

$$y - \bar{Y} = \beta_1 (x_1 - \bar{X}_1) + \beta_2 (x_2 - \bar{X}_2).$$

Y recopilando todos los valores obtenidos, tenemos que

$$y - \bar{Y} = \frac{S^2_{X_2} S_{X_1 Y} - S_{X_1 X_2} S_{X_2 Y}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1 X_2}} (x_1 - \bar{X}_1) + \frac{S^2_{X_1} S_{X_2 Y} - S_{X_1 X_2} S_{X_1 Y}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1 X_2}} (x_2 - \bar{X}_2).$$

Es evidente, que la expresión del denominador de β_1 y β_2 no debe anularse para poder obtener una solución de estos parámetros. Es decir,

$$S^2_{X_1} S^2_{X_2} - S^2_{X_1 X_2} \neq 0.$$

Centro de gravedad

El centro de gravedad de una nube de puntos es aquel punto formado por las medias de todas las variables (endógena y exógenas) que construyen el hiperplano de regresión. En el caso de tres variables, el centro de gravedad es:

$$(\bar{X}_1, \bar{X}_2, \bar{Y})$$

Supongamos que las medias de las variables Y, X_1 y X_2 son: 82,59, 0,34 y 523,59. Entonces, estos valores serán el centro de gravedad de la nube de puntos entre estas tres variables.

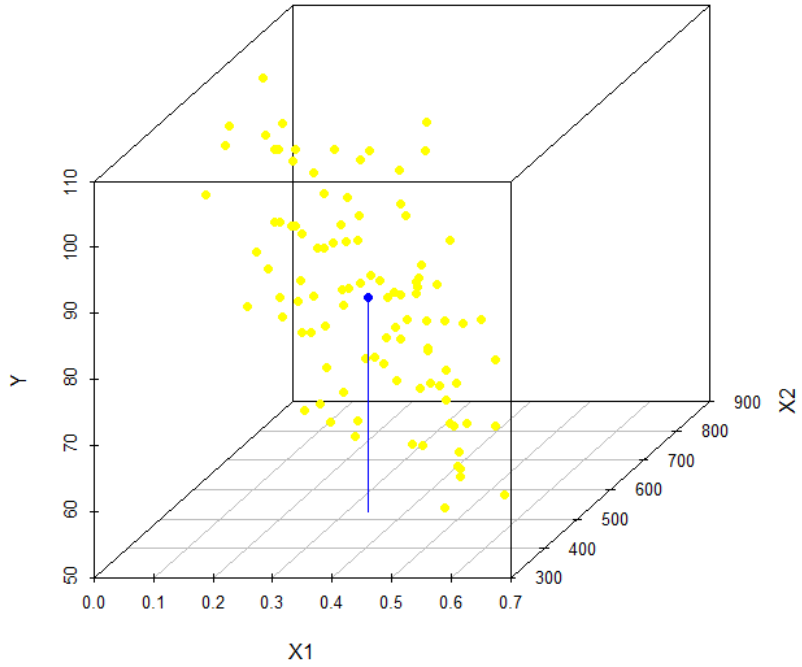


Figura 20.6: Representación del centro de gravedad del plano de regresión de Y sobre X_1 y X_2

Planos de regresión

El plano que hemos obtenido se conoce como *plano de regresión* de Y sobre X_1 y X_2. También es posible obtener los planos de regresión de X_1 sobre X_2 e Y, y el de X_2 sobre X_1 e Y. Estos planos se pueden expresar como:

$$x_1 = \beta_0' + \beta_1' x_2 + \beta_2' y,$$
$$x_2 = \beta_0'' + \beta_1'' x_1 + \beta_2'' y,$$

Supongamos que deseamos analizar la relación entre las variables peso (Y), consumo de sustancias adictivas (X_1) e ingesta de grasas (X_2). Gráficamente, podríamos realizar las siguientes representaciones en relación con qué variable consideramos como endógena:

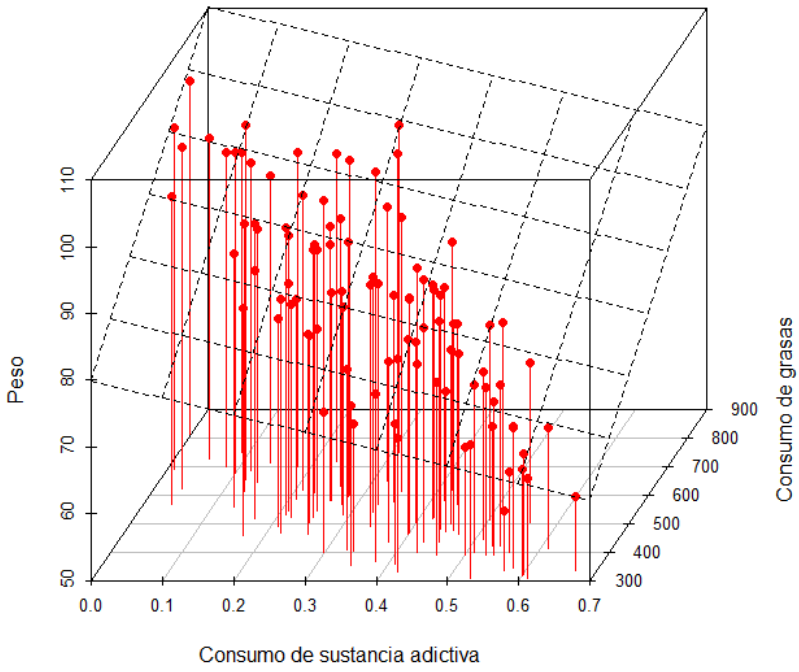


Figura 20.7: Representación del plano de regresión de Y sobre X_1 y X_2

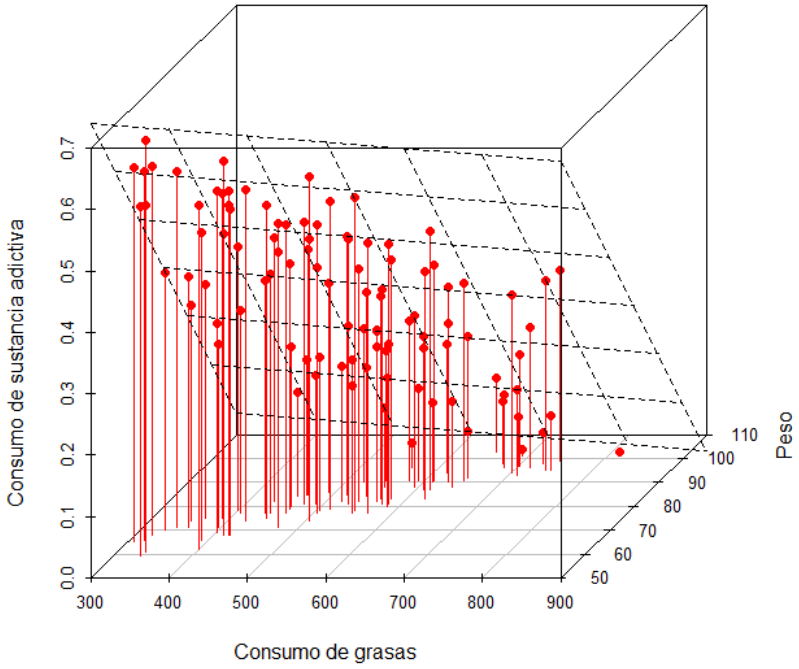


Figura 20.8: Representación del plano de regresión de X_1 sobre X_2 e Y

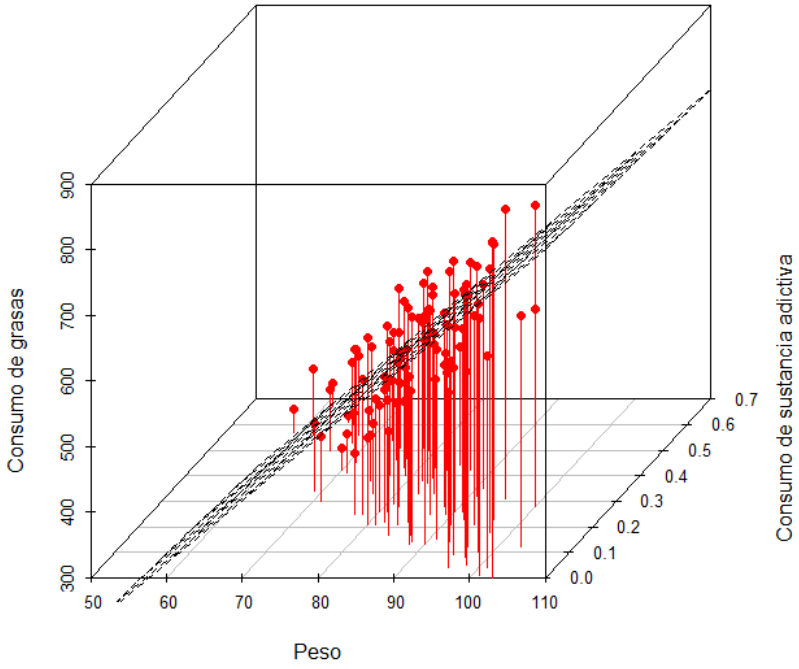


Figura 20.9: Representación del plano de regresión de X_2 sobre Y y X_1

Predicción

Ante una nueva observación, la estimación del valor de y_{N+1} se realiza sustituyendo en la ecuación del hiperplano los valores: x_{1N+1}, x_{2N+1}, ..., x_{kN+1}. Es decir, se sustituyen los valores nuevos en la siguiente expresión:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$

Ejemplo 3. Sean las variables Y, X_1 y X_2 de las que se disponen los datos siguientes procedentes de un laboratorio:

Y	X_1	X_2
1	40	2
2	50	3
3	70	4
4	75	6
5	80	7
6	100	8

Se desea determinar el plano de regresión de Y sobre X_1 y X_2 para estimar el valor de y cuando $x_1 = 68$ y $x_2 = 5$. Esto debe hacerse por mínimos cuadrados.

En primer lugar, podemos representar la nube de puntos de nuestros datos.

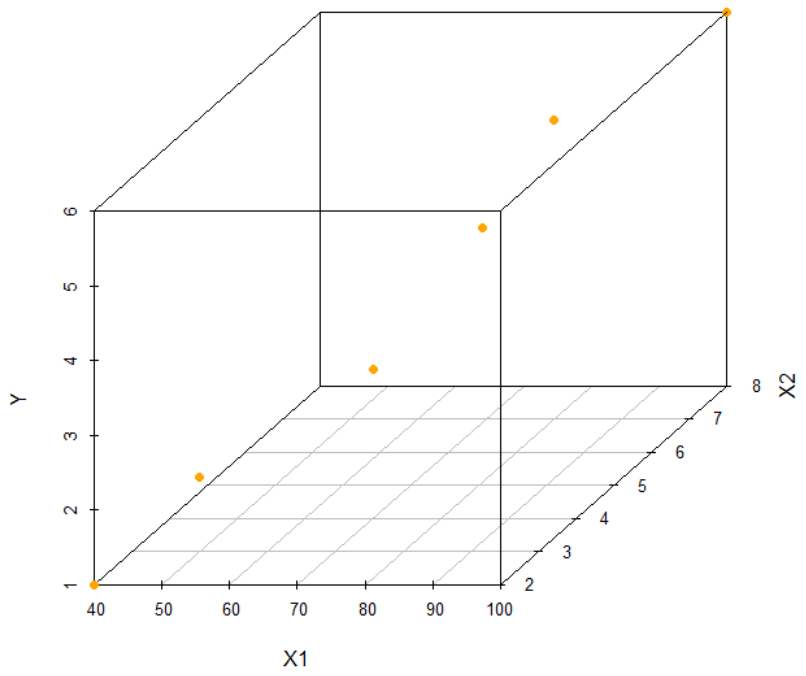


Figura 20.10: Representación de la nube de puntos

Para construir el plano de regresión, vamos a crear las siguientes columnas:

y_i^2	x_{1i}^2	x_{2i}^2	$y_i x_{1i}$	$y_i x_{2i}$	$x_{1i} x_{2i}$
1	1600	4	40	2	80
4	2500	9	100	6	150
9	4900	16	210	12	280
16	5625	36	300	24	450
25	6400	49	400	35	560
36	10000	64	600	48	800
$\sum = 91$	$\sum = 31025$	$\sum = 178$	$\sum = 1650$	$\sum = 127$	$\sum = 2320$

El vector de medias es:

$$\begin{pmatrix} \bar{Y} \\ \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \begin{pmatrix} 3,50 \\ 69,17 \\ 5,00 \end{pmatrix}$$

La representación del centro de gravedad sería:

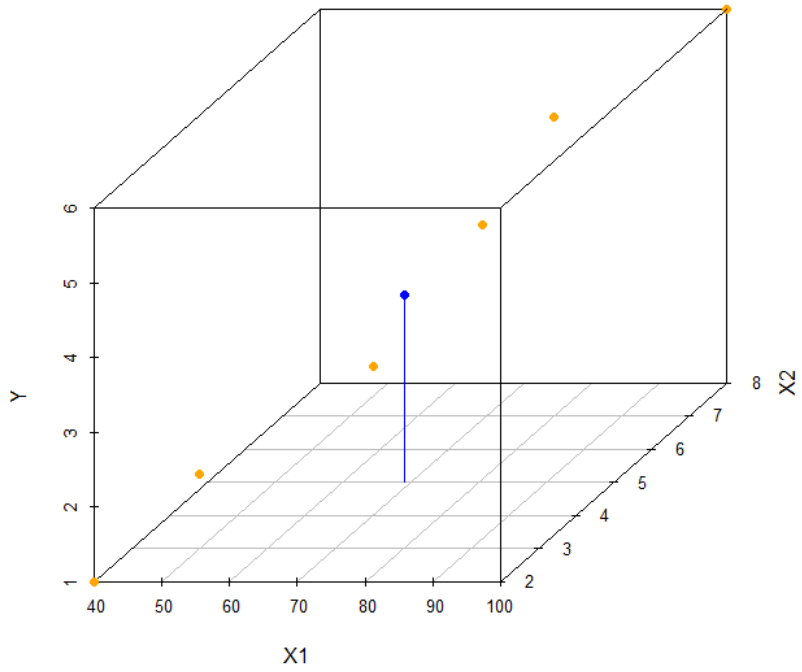


Figura 20.11: Representación del centro de gravedad del plano de regresión de Y sobre X_1 y X_2

La matriz de varianzas-covarianzas es:

$$S_{YX_1X_2} = \begin{pmatrix} 2,92 & 32,92 & 3,67 \\ 32,92 & 386,81 & 40,83 \\ 3,67 & 40,83 & 4,67 \end{pmatrix}$$

Ahora nuestro objetivo es determinar los valores de β_0, β_1 y β_2 para hallar el plano de regresión $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Sustituimos:

$$\beta_1 = \frac{S^2_{X_2} S_{X_1 Y} - S_{X_1 X_2} S_{X_2 Y}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1 X_2}} = 0,0282$$
$$\beta_2 = \frac{S^2_{X_1} S_{X_2 Y} - S_{X_1 X_2} S_{X_1 Y}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1 X_2}} = 0,5387$$
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 = -1,1462$$

Y nuestro plano es:

$$y = -1,1462 + 0,0282 x_1 + 0,5387 x_2.$$

La representación del plano de regresión de Y sobre X_1 y X_2:

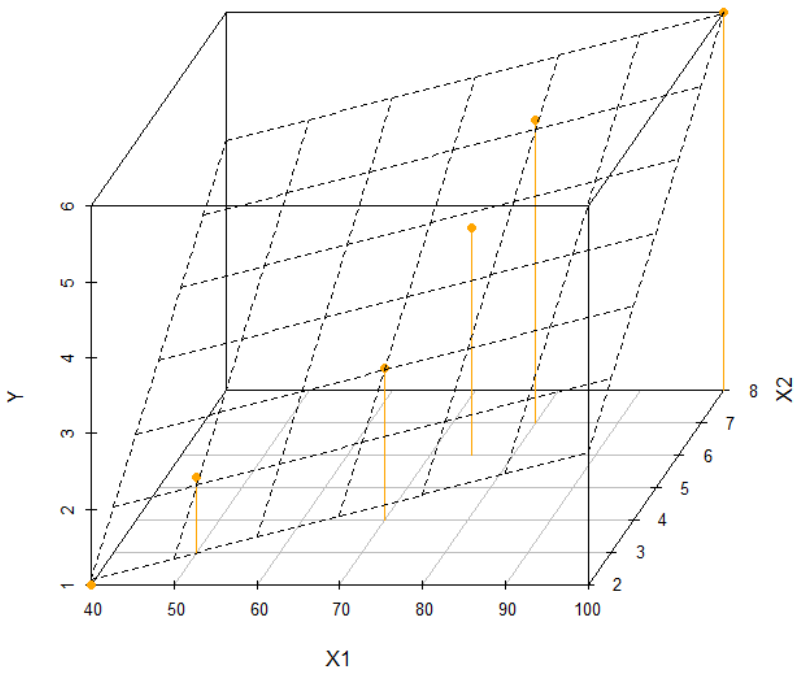


Figura 20.12: Representación del plano de regresión de Y sobre X_1 y X_2

Por último, la predicción es:

$$\hat{y} = -1,1462 + 0,0282 * 68 + 0,5387 * 5 = 3,4649.$$

20.3.3 Análisis de los residuos: correlación múltiple y parcial

La correlación múltiple se analiza de forma análoga al caso bidimensional.

Varianza residual

La varianza residual se obtiene como

$$S_e^2 = \frac{\sum_{i=1}^{N} e_i^2}{N},$$

donde:
$e_i = y_i - \hat{y}_i$, lo que equivale a que $y_i = \hat{y}_i + e_i$.

La relación entre la varianza residual y la varianza de Y es:

$$S_Y^2 = S_{\hat{Y}}^2 + S_e^2.$$

El coeficiente de determinación múltiple

El coeficiente de determinación múltiple toma valores entre 0 y 1 y mide en qué grado la variable Y depende de las variables exógenas X_1, X_2, ..., X_k. Este coeficiente tiene en cuenta las variables exógenas de forma conjunta y se suele expresar en porcentajes. Se denota por R^2 (como en el caso bidimensional) y viene expresado como:

$$R^2 = \frac{S_{\hat{Y}}^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2}.$$

- Valores de R^2 próximos a 0 indican un ajuste inadecuado y que las variables exógenas no explican la variable endógena (es decir, independencia entre ellas). Si el valor es 0, el ajuste indica que todas las variaciones son debidas a los errores.

  En la siguiente gráfica se muestra un ejemplo de un caso donde el valor del coeficiente de determinación múltiple es próximo a cero.

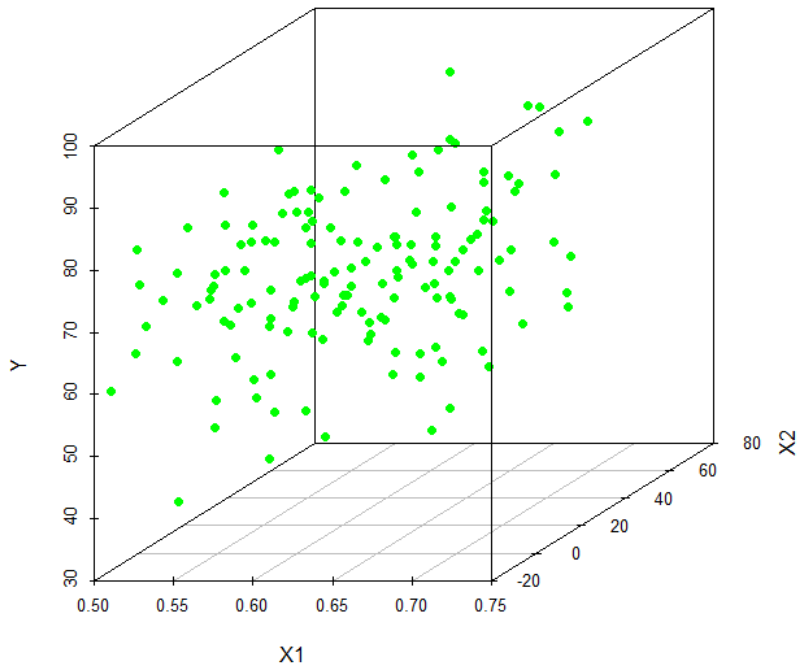


Figura 20.13: Representación de la relación de las variables Y, X_1 y X_2: coeficiente de determinación múltiple = 0,0107

- Valores de R^2 próximos a 1 indican un ajuste aceptable. Es decir, cuanto más alto sea el valor de R^2, mejor ajuste y mayor proporción de variabilidad que se consigue explicar con nuestro modelo. Si el valor es 1, el ajuste es perfecto.

  En la siguiente gráfica se muestra un ejemplo de un caso donde el valor del coeficiente de determinación múltiple toma un valor alto.

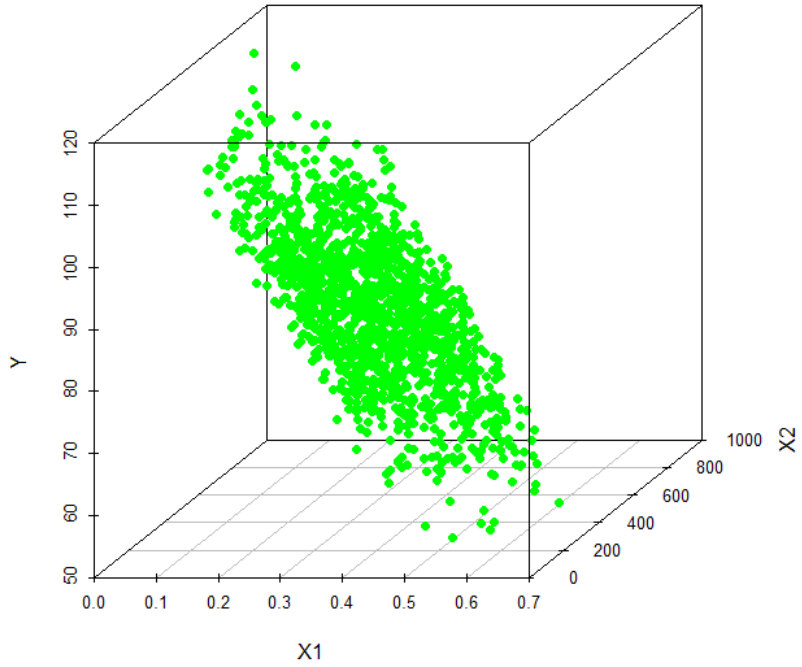


Figura 20.14: Representación de la relación de las variables Y, X_1 y X_2: coeficiente de determinación múltiple = 0,8317

El coeficiente de correlación lineal múltiple

El coeficiente de correlación lineal múltiple de Y respecto de las variables X_1, X_2, ..., X_k se obtiene como la raíz cuadrada positiva del coeficiente de determinación múltiple. También toma valores entre 0 y 1. Se denota por r y viene expresado como:

$$r = +\sqrt{R^2}.$$

El coeficiente de correlación parcial

Cuando nos quedemos con un subconjunto de las variables exógenas, tendremos los coeficientes de correlación parcial. Si nos quedamos con dos o más variables exógenas, el coeficiente de correlación parcial se obtiene como lo hemos visto anteriormente. Pero si nos quedamos con una única variable exógena (o incluso si se considera la regresión de Y sobre cada una de las variables exógenas), el coeficiente de correlación parcial se obtiene como en el caso del modelo de regresión lineal simple. La fórmula es la siguiente:

$$r = \frac{S_{YX}}{S_X S_Y}.$$

No obstante, otro modo más habitual de medir la correlación parcial de Y y X_1, manteniendo X_2 constante viene expresada por la siguiente expresión:

$$r_{YX_1.X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{YX_2}}\sqrt{1 - r^2_{X_1X_2}}}.$$

r representa el coeficiente de correlación de Pearson. Y la expresión anterior se puede extraplorar para cualquier combinación de variables exógenas. Cuando no existe correlación entre la variable Y y la variable exógena X_2 y tampoco entre X_1 y X_2, el coeficiente de correlación parcial anterior coincidirá con el coeficiente de correlación de Person (que es el caso expuesto al principio).

En la práctica, usaremos este último coeficiente de correlación parcial que mide la variación conjunta entre dos variables (una dependiente y otra independiente) controlando por el efecto de una tercera variable (independiente).

Ejemplo 4. Retomando los datos del Ejemplo 3, vamos a calcular el coeficiente de determinación múltiple, el coeficiente de correlación múltiple y los coeficientes de correlación parciales.

Empezamos construyendo las siguientes columnas:

y_i	x_{1i}	x_{2i}	$\hat{y}_i$	e_i	e_i^2
1	40	2	1.0605	-0.0605	0.0037
2	50	3	1.8815	0.1185	0.0140
3	70	4	2.9849	0.0151	0.0002
4	75	6	4.2034	-0.2034	0.0414
5	80	7	4.8832	0.1168	0.0136
6	100	8	5.9866	0.0134	0.0002

Por tanto, el coeficiente de determinación múltiple es:

$$R^2 = \frac{S^2_{\hat{Y}}}{S^2_Y} = 1 - \frac{S^2_e}{S^2_Y} = 1 - \frac{0,0122}{2,92} = 0,9958.$$

En nuestro caso, el R^2 toma un valor muy alto cercano a 1. Por lo que podemos concluir que hemos conseguido explicar el 99,58 % de la variabilidad de nuestro problema.

Por su parte, el coeficiente de correlación múltiple es:

$$r = +\sqrt{R^2} = 0,9979.$$

Este coeficiente también presenta un valor alto. Lo cual contribuye a determinar que existe una alta dependencia entre la variable Y y las variables exógenas X_1 y X_2 en conjunto.

Por otro lado, también es posible calcular el coeficiente de correlación lineal entre Y y X_1, y entre Y y X_2.

$$r_{YX_1} = \frac{32,92}{\sqrt{2,92 * 386,81}} = 0,9795$$

$$r_{YX_2} = \frac{3,67}{\sqrt{2,92 * 4,67}} = 0,9938$$

El coeficiente de correlación parcial de Y y X_1, manteniendo X_2 constante es:

$$r_{YX_1.X_2} = 0,8021.$$

Previamente, hemos obtenido el coeficiente de correlación entre las variables exógenas (0,9607).

20.4 Multicolinealidad

La *multicolinealidad* es un problema que se produce cuando se desean calcular los coeficientes de la ecuación de regresión y nos encontramos que al menos dos variables exógenas son linealmente dependientes. Este hecho produce que el sistema de ecuaciones que pretendemos resolver carezca de solución, dado que existe proporcionalidad entre variables (por ejemplo, entre x_{1i} y x_{2i}). A nivel matricial, esto supone que el determinante de la matriz que queremos invertir es cero. Lo que la hace invertible.

En el caso del modelo de regresión lineal para 3 variables, si existe multicolinealidad (o colinealidad), y la correlación entre las variables exógenas es perfecta, pasará que $S^2_{X_1}S^2_{X_2} - S^2_{X_1X_2} = 0$. Por tanto, no se pueden obtener los valores de β_0, β_1 y β_2.

En el modelo de regresión lineal múltiple, para solucionar este problema se puede seleccionar previamente un subconjunto de variables exógenas entre las cuales no exista una alta o perfecta correlación (o dependencia) entre ellas y expliquen mejor la variabilidad de la variable endógena.

Ejemplo 5. Retomando los datos del Ejemplo 3, vamos a analizar la relación entre las variables exógenas del modelo.

En primer lugar, el coeficiente de correlación entre las variables exógenas X_1 y X_2 toma un valor muy alto:

$$r_{X_1X_2} = \frac{40,83}{\sqrt{386,81 * 4,67}} = 0,9607.$$

En la nube de puntos también se aprecia la relación positiva entre ambas variables:

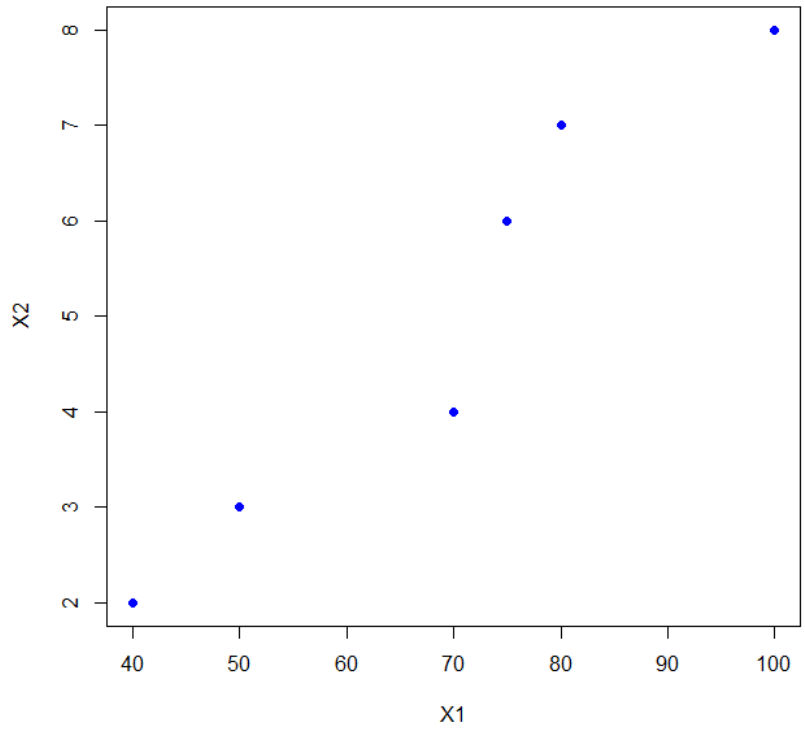


Figura 20.15: Representación de la nube de puntos de X_1 y X_2

En segundo lugar, comprobamos que $S^2_{X_1}S^2_{X_2} - S^2_{X_1X_2}$ no sea próximo o igual cero.

$$S^2_{X_1}S^2_{X_2} - S^2_{X_1X_2} = 137,7315.$$

Por último, si consideramos que existe una alta correlación entre las variables X_1 y X_2, podemos quedarnos con la variable X_2 para explicar la variable Y. Recordemos que el coeficiente de correlación era superior con la variable X_2 ($r_{YX_2} = 0,9938$, frente a $r_{YX_1} = 0,9795$).

En la práctica se utilizan el factor de inflación de la varianza (FIV) y la tolerancia (T) para analizar la presencia de multicolinealidad en un modelo de regresión lineal múltiple. Estos indicadores son fáciles de obtener en la mayoría de paquetes estadísticos.

El factor de inflación de la varianza (FIV) se calcula como:

$$FIV = \frac{1}{1 - R_j^2},$$

donde:
R_j^2 es el coeficiente de determinación de la regresión de la variable X_j sobre el resto de variables exógenas.

FIV toma valores positivos.

- Si FIV = 1, no hay problemas de multicolinealidad entre la variable exógena estudiada con el resto de variables exógenas.
- Si FIV se encuentra entre 1 y 5, se considera que existe un problema de multicolinealidad bajo entre la variable exógena estudiada con el resto de variables exógenas.
- Si FIV se encuentra entre 5 y 10, se considera que existe un problema de multicolinealidad medianamente alto entre la variable exógena estudiada con el resto de variables exógenas.
- Si FIV >10, se considera que existe un problema serio de multicolinealidad entre la variable exógena estudiada con el resto de variables exógenas.

La tolerancia (T) es el denominador del factor de inflación de la varianza. Es decir,

$$T = 1 - R_j^2.$$

Se considera que existen serios problemas de multicolinealidad cuando la tolerancia es inferior a 0,10.

Por otro lado, también se puede entender que tenemos serios problemas de multicolinealidad cuando $R_j^2 > 0,9$.

Retomando nuestro ejemplo, los valores de FIV y T son:

$$FIV = \frac{1}{1 - R_j^2} = \frac{1}{1 - 0,9607^2} = 12,9777$$

$$T = 1 - R_j^2 = 1 - 0,9607^2 = 0,0771$$

A la vista de los valores del FIV y T, podemos concluir que tenemos serios problemas de multicolinealidad entre las variables exógenas X_1 y X_2.

En nuestro ejemplo sólo tenemos dos variables exógenas, por lo que FIV y T toman el mismo valor para las dos rectas de regresión que se pueden crear con las variables X_1 y X_2.

Problema propuesto. Se desea calcular el plano de regresión de Y sobre X y Z para predecir una nueva observación que toma los valores $x = 500,671$ y $z = 5$. Los datos se presentan en la siguiente tabla:

Y	X	Z
296	466.715	5
211	443.85	7
148	474.785	6
247	800.275	4
48	626.77	5
332	496.305	1
439	446.54	7
88	637.53	4
78	687.295	3
373	305.315	5
245	317.42	7
456	613.32	3
472	473.44	5
5	801.62	4
41	689.985	6
131	232.685	6
324	203.095	3
429	396.775	3

Por último, se recomienda la lectura de otros libros de varios autores para ampliar los conocimientos respecto de la materia estudiada (Tomeo Perucha y Uña Juárez 2009; Peña 2001).

Bibliografía

Peña, Daniel (2001). *Fundamentos de estadística*. Madrid: Alianza editorial.

Tomeo Perucha, Venancio e Isaías Uña Juárez (2009). *Estadística descriptiva*. Madrid: Ibergarceta Publicaciones.

Montero Granados, Roberto (2016). "Modelos de regresión lineal múltiple". En: *Granada, España: Departamento de Economía Aplicada, Universidad de Granada* (página 86).

Tema 21

Números índices. Los índices simples. Propiedades de los índices simples. Índices complejos. Índices de Laspeyres y Paasche. Índices de precios, de volumen y de valor.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

AM Montiel Torres, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson

P. Martín Guzmán (2006). *Manual de estadística: descriptiva*. Thomson - Civitas

José Miguel Casas Sánchez y col. (2010). *Estadística para las ciencias sociales*. Editorial Universitaria Ramón Areces

21.1 Introducción

En los temas anteriores se han analizado variables estadísticas aisladas desde el punto de vista descriptivo y la relación entre ellas a través de la correlación y de la regresión. Para su estudio no se ha tenido en cuenta el momento de recogida de los datos.

En este tema se van a analizar las variables estadísticas a lo largo del tiempo mediante la creación de distintos indicadores denominados **números índices**. Estos índices nos ayudan a abordar el problema de la comparación de una serie de observaciones respecto a una situación inicial.

Lo normal es que estas comparaciones se hagan mediante cocientes para eliminar las unidades de la variable que se está midiendo.

Sectores como la economía y el mundo de los negocios se caracterizan por su dinamismo. Sin embargo, es necesario realizar comparaciones que indiquen su evolución, tanto en el tiempo como por zonas geográficas, y para ello se necesitan medidas estadísticas referenciadas a un momento o una zona determinada.

número índice: un número índice es una medida estadística que permite estudiar la

evolución que se produce en un magnitud simple o compleja con respecto al tiempo o al espacio.

Los números índices están incorporados en nuestra vida diaria. Un ejemplo es el Índices de Precios de Consumo de España que se utiliza constantemente para actualizar pensiones, alquileres, etc.

Para crear distintos números índices debemos definir:

periodo base o de referencia este momento se fija arbitrariamente por el investigador. Elegir el periodo base es uno de los principales problemas pues el periodo de referencia debe ser un momento donde ninguna de las variables que vayamos a medir sea atípica. Por ejemplo este año 2021 o el año pasado no deberán ser periodos de referencia en economía, debido a la pandemaia de la COVID-19.

periodo actual que es el momento que se desea comparar con el periodo de referencia.

21.2 Índices simples

Denominamos índices simples a aquellos que están referidos a una única magnitud, miden la variación en tanto por uno entre el dato de referencia y el dato actual. Al periodo o la zona que se elige de referencia se le asigna el valor 100, el resto de los valores nos indicarán el porcentaje de variación que se ha producido respecto al periodo o zona base.

Es muy habitual multiplicar por cien el número índice y expresarlo en porcentaje.

Los índices simples más utilizados son:

Precio relativo: es la razón entre el precio de un bien en el momento actual y el precio en el perido de referencia.

Cantidad relativa: es la razón entre la cantidad producida o consumida de un cierto producto en el momento actual (zona estudiada) y la cantidad producida o consumida en el periodo de referencia (zona de referencia).

Valor relativo: el valor de un bien se calcula como el producto de la cantidad del bien por su precio ($P_i * Q_i$). El valor relativo será el cociente del valor actual entre el valor de referencia.

Veamos un ejemplo donde se analiza la evolución del precio de un determinado producto.

Año	Precio del bien
2010	1,2
2011	1,5
2012	1,8
2013	2,1
2014	2,2
2015	2,4

Tabla 21.1: Año y precio del bien

Se define el número índice simple como el cociente entre el precio en el momento actual y su precio en el momento de referencia.

$$I_0^t = \frac{Y_t}{Y_0} * 100$$

Siendo 0 el periodo de referencia y t el momento que se está valorando.

Año	Índice de precio base 2010
2010	100 %
2011	125 %
2012	150 %
2013	175 %
2014	183,3 %
2015	200 %

Tabla 21.2: Índice de precios de un bien con referencia a 2010

La interpretación de este índice de precios es muy sencilla. Se observa que en el año de referencia (2010) el valor es el 100 % y en 2015 es el 200 %, es decir entre el 2010 y el 2015 el valor del bien se ha duplicado.

Si se quiere comparar entre dos o más productos, cuál de ellos ha sido el que más se ha revalorizado se deben utilizar los números índices porque al no tener unidades no depende del precio del bien.

Año	Precio del artículo 1	Precio del artículo 2
2010	1,2	200
2011	1,5	250
2012	1,8	275
2013	2,1	300
2014	2,2	325
2015	2,4	375

Tabla 21.3: Precio medio de dos artículos desde 2010 a 2015

Calculamos los números índices para cada artículo y obtenemos los siguientes resultados:

Año	Índice artículo 1	Índice artículo 2
2010	100	100 %
2011	125	125 %
2012	150	137,5 %
2013	175	150 %
2014	183,3	162,5 %
2015	200	187,5 %

Tabla 21.4: Índice de precios de los artículos 1 y 2 con referencia a 2010

En la Tabla 21.4 se observa que el crecimiento del precio del artículo 1 fue mayor para cualquier año que el del artículo 2.

A partir de los números indices se puede calcular la tasa de variación porcentual.

Si Y_{t_1} es el valor de un bien en un periodo t_1 y Y_{t_2} el valor en un periodo t_2. A partir del número índice $I_{t_1}^{t_2}$ podemos obtener su tasa de variación.

Se define como **tasa de variación o variación porcentual** al cociente entre el incremento del valor del bien en el periodo $[t_1, t_2]$ y su valor en el instante t_1.

$$\begin{aligned} Ta_{t_1}^{t_2}(Y) &= \frac{Y_{t_2}-Y_{t_1}}{Y_{t_1}} * 100 = \\ &= \left(\frac{Y_{t_2}}{Y_{t_1}} - 1\right) * 100 = \\ &= I_{t_1}^{t_2} - 100 \end{aligned}$$

Tabla 21.5: Valores de las n variables en k periodos

Una de las tasas más utilizadas es la **tasa de variación interanual**, cuando las observaciones correspenden al valor del bien en dos años consecutivos.

21.3 Propiedades de los índices simples

Es importante enumerar algunas de las propiedades de los números índices simples.

Existencia: el número índice siempre existe y es finito.

Identidad: siempre

$$I_t^t = 100\,\%$$

Propiedad circular: esta propiedad es interesante para cuando es necesario cambiar el periodo de referencia.

$$I_0^t = \frac{Y_t}{Y_0} * 100 = I_{t_1}^t * I_0^{t_1}$$

Inversión:

$$I_0^t = \frac{1}{I_t^0}$$

Encadenamiento:

$$I_0^t = I_{t-1}^t * I_{t-2}^{t-1} * I_{t-3}^{t-2} * ... * I_0^1$$

Esta propiedad es muy utilizada debido a la complejidad que tiene elegir el periodo de referencia y la pérdida de información del índice a medida que transcurre el tiempo. Parece mucho más importante comparar los precios de la luz actuales con los del año pasado que con los que tenía hace 20.

Adición: si tenemos dos magnitudes simples X y Y y $V = X + Y$ entones

$$I_0^t(V) = \frac{X_t + Y_t}{X_0 + Y_0} * 100 = I_0^t(X)\frac{X_0}{X_0 + Y_0} + I_0^t(Y)\frac{X_0}{X_0 + Y_0}$$

Multiplicación: si tenemos dos magnitudes simples P y Q y $V = P * Q$ entones

$$I_0^t(V) = \frac{V_t}{V_0} * 100 = I_0^t(P)I_0^t(Q)$$

Homogeneidad: si realizamos una transformación $Y = aX$ el índice queda invariante.

21.4 Índices complejos

En la sección anterior se trabajó con números índices simples, sin embargo donde es interesante utilizar números índices es en situaciones más complejas donde intervienen muchas variables.

Índice complejo: llamaremos índice complejo a un único índice que sintetiza la información suministrada por los distintos índices simples para cada una de las variables. Para resumir la información de los índices simples se puede utilizar la media aritmética, la geométrica, la agregativa o la armónica.

Un ejemplo de un índice compuesto es el Índice de Precios al Consumo (IPC), este índice hace referencia al nivel de precios de los artículos de una determinada cesta elaborada con los productos más consumidos.

Los índices complejos pueden ser ponderados o no ponderados.

21.5 Principales índices no ponderados

Índice no ponderado:

Consideremos que tenemos varios bienes $(Y_1, Y_2, ..., Y_n)$ a lo largo de k periodos. A partir de los números índices simples se calcula el número índice complejo.

Periodo	Artículo 1	Artículo 2	...	Artículo n
1	y_{11}	y_{12}	...	y_{1n}
2	y_{21}	y_{22}	...	y_{2n}
3	y_{31}	y_{32}	...	y_{3n}
...	...	...	...	...
k	y_{k1}	y_{k2}	...	y_{kn}

Tabla 21.6: Valores de las n variables en k periodos

21.5.1 Índice de Bradstreet y Dûtot:

se calcula mediante la media agregativa de las n variables.

$$I_1^i = \frac{\sum_{j=1}^{n} y_{ij}}{\sum_{j=1}^{n} y_{1j}}$$

Periodo	Precio art. 1	Precio art. 2	Precio art. 3	Índice
2012	125	25	0,25	100 %
2013	150	30	0,30	120 %
2014	160	32	0,31	127,99 %
2015	175	40	0,50	143,43 %

Tabla 21.7: Índice de Bradstreet y Dûtot

Este método solo se utiliza cuando todas las variables están medidas en las mismas unidades.

21.5.2 Índice de Sauerbek:

Si las variables no vienen medidas en las mismas unidades, el procedimiento que se sigue es calcular el índice simple para cada variable y después hallar la media aritmética de los distintos índices simples.

$$I_0^t = \frac{\sum_{j=1}^{n} I_0^t(Y_j)}{n}$$

Periodo	Naranjas (kg.)	Sal (gr.)	Aceite (litros)
2012	125	250	100
2013	150	130	90
2014	160	320	120
2015	175	400	150

Tabla 21.8: Valores de las variables en los distintos periodos

Se calculan los índices simples para naranjas, sal y aceite.

Periodo	Naranjas (kg.)	Sal (gr.)	Aceite (litros)	Índice complejo
2012	100 %	100 %	100 %	100 %
2013	120 %	52 %	90 %	87,33 %
2014	128	128 %	120 %	125,33 %
2015	140 %	160 %	150 %	150 %

Tabla 21.9: Índice de Sauerbek

21.6 Principales índices complejos ponderados

Índice ponderado:

Los índices no ponderados presentan el inconveniente de considerar a todas las variables igual de importantes. Cuando no todas las variables pesan lo mismo se debe asignar a cada una ellas una ponderación w_i. Se define el índice complejo ponderado como la media ponderada de los índices simples.

Si las magnitudes de las variables son homogéneas, se utiliza la media agregativa ponderada.

$$I_0^i = \frac{\sum_{j=1}^{n} y_{ij} w_j}{\sum_{j=1}^{n} y_{0j} w_j} * 100\,\%$$

Si las variables no están medidas en las mismas unidades, entonces se calcula la media ponderada de los índices simples.

$$I_0^t = \frac{\sum_{j=1}^{n} I_0^t(Y_j) w_j}{\sum_{j=1}^{n} w_j} * 100\,\%$$

Un caso particular de los índices complejos sería la evolución del valor de los productos de un determinado almacén o de una determinada cesta de alimentos.

De forma general, se expresa por p_{i0} y q_{i0} los precios y las cantidades de los productos que intervienen en la cesta para la cual se va a calcular su valor en el periodo de referencia y p_{it} y q_{it} los precios y las cantidades en el momento actual. A partir de estos datos se formula el índice general del valor como:

$$I_0^t = \frac{\sum_{i=1}^{n} p_{it} q_{it}}{\sum_{i=1}^{n} p_{i0} q_{i0}} * 100$$

Siendo n el número de productos.

Supongamos ahora que se desea saber que parte de la variación del valor de la cesta de productos es debido a la variación de los precios y que parte es debido a la variación de las cantidades.

21.6.1 Índice de Laspeyres

Índice de precios de Laspeyres

Se define **el índice de precios de Laspeyres (IPL)** como un índice ponderado que estudia la variación del valor de una cesta de productos suponiendo que las cantidades permanecen constantes en el tiempo, es decir, debida a la variación de los precios.

$$IPL_0^t = \frac{\sum_{i=1}^{n} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} * 100$$

Veamos un ejemplo del índice de precios Laspeyres

Primero calculamos los índices simples de precios para cada producto:

Producto	precio de referencia	precio actual	Índice simple
Prod. 1	0,5	0,4	80 %
Prod. 2	1,2	1,5	125 %
Prod. 3	3,5	3,8	108,57 %

Tabla 21.10: Índice simple de los precios de una cesta de productos

Se observa que el único producto que bajo su precio fue el Prod. 1 ($80\,\%$) y el que más subió fue el Prod. 2 ($125\,\%$). A partir de estos índices simples lo que se calcula es la variación del valor de la cesta formada por los tres productos pero cada uno con un

peso diferente (el peso es proporcional a la cantidad de cada producto que hay en la cesta)

Producto	p_{i0}	q_{i0}	$p_{i0}q_{i0}$	Índice simple	$I_0^t(Prod.i) * p_{i0}q_{i0}$
Prod. 1	0,5	3	1,5	80 %	120
Prod. 2	1,2	5	6	125 %	750
Prod. 3	3,5	8	28	108,57 %	3040
Total			35,5		3910

Tabla 21.11: Ponderación de los precios de los productos de la cesta por las cantidades

Otra forma de definir el índice de precios de Laspeyres es:

$$IPL_0^t = \frac{\sum_{i=1}^{n} I_0^t(Prod.i) * p_{i0}q_{i0}}{\sum_{i=1}^{N} p_{i0}q_{i0}} * 100$$

A partir de la tabla 21.10 se obtiene el índice de Laspeyres como

$$\frac{3910}{35,5} = 110,14$$

Índice de cantidades de Laspeyres

En economía existe una gran variedad de cantidades, pero las más importantes son los volúmenes producidos por las empresas o los consumidos por las familias.

El índice de cantidades de Laspeyres (IQL) valora la variación del valor de una producción o de un consumo debida a la modificación de las cantidades producidas o consumidas. Las cantidades q_{it} son ponderadas por el precio final de venta si nos referimos a consumo o valor añadido por unidad producida si nos referimos a producción p_{i0}.

$$IQL_0^t = \frac{\sum_{i=1}^{n} q_{it}p_{i0}}{\sum_{i=1}^{N} q_{i0}p_{i0}} * 100$$

Ejemplos de índices de Laspeyres más utilizados en la economía española son:

Índice de Producción Industrial para su obtención se realiza una encuesta de periodicidad mensual que sondea todos los meses más de 9.000 establecimientos. Este indicador mide la evolución de la producción industrial excluida la construcción.

Índice de Precios Industriales este indicador mide la evolución de los precios de los productos fabricados y vendidos en el mercado interior. Para su calculo se realiza mensualmente una encuesta en más de 6.000 establecimientos industriales.

21.6.2 Índice de Paasche

Índice de precios de Paasche

Si se analiza el índice de precios de Laspeyres parece que no es muy lógico pensar que las cantidades de cada producto van a mantenerse constantes en el tiempo. Por ejemplo no parece lógico pensar que la cantidad de productos infantiles que consume una familia no van a variar a lo largo del tiempo.
Es este el motivo que ha llevado a los economistas a diseñar otros índices que midan la variación del valor de la cesta debido a los precios pero sin la restricción de que todos los productos se ponderen por las mismas cantidades a lo largo del tiempo.
El índice de precios de Paasche (IPP) se define como

$$IPP_0^t = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{N} p_{i0}q_{it}} * 100$$

Es decir, se mide la variación del valor de la cesta debido a la variación de los precios suponiendo que en el periodo de referencia se adquirieron las cantidades las mismas cantidades que en el año actual.

Veamos cual es el índice de Paasche para el ejemplo anterior:

Producto	p_{i0}	p_{it}	q_{it}	$p_{i0}q_{it}$	$p_{it}q_{it}$
Prod. 1	0,5	0,4	4	2	1,6
Prod. 2	1,2	1,5	6	7,2	9
Prod. 3	3,5	3,8	8	28	30,4
Total				37,2	41

Tabla 21.12: Ponderación de los precios de los productos de la cesta por las cantidades en el año actual

En este caso el índice de precios según Paasche vendrá dado por el cociente

$$\frac{41}{37,2} * 100 = 110,22$$

El índice de Paasche indica que la subida ha sido no de un $10,14\,\%$ como indicaba el índice de Laspeyres, si no de un $10,22\,\%$, la diferencia es que las cantidades de los productos de la cesta habían variado del periodo 0 al periode t.

Índice de cantidades Paasche

En este caso los coeficientes de ponderación son medidos como en el caso del índice de precios de Paasche en el momento actual y no en el periodo de referencia. Así el índice de cantidades Paasche (IQP) se define como:

$$IQP_0^t = \frac{\sum_{i=1}^{n} q_{it}p_{it}}{\sum_{i=1}^{N} q_{i0}p_{it}} * 100$$

21.6.3 Índice de Edgeworth

Índice de precios de Edgeworth

En este índice que denotaremos como IPE se utilizan como coeficientes de ponderación la suma de los utilizados en los índices de Laspeyre y Paasche.

$$w_i = p_{i0}q_{i0} + p_{i0}q_{it}$$

Obteniendo la expresión:

$$IPE_0^t = \frac{\sum_{i=1}^{n} \frac{p_{it}}{p_{i0}} w_i}{\sum_{i=1}^{n} w_i} * 100 =$$

$$IPE_0^t = \frac{\sum_{i=1}^{n} p_{it}(q_{i0} + q_{it})}{\sum_{i=1}^{n} p_{i0}(q_{i0} + q_{it})} * 100 =$$

$$IPE_0^t = \frac{\sum_{i=1}^{n} p_{it}q_{i0} + \sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{i0} + \sum_{i=1}^{n} p_{i0}q_{it}} * 100$$

Si utilizamos los datos de la Tabla 21.10 y 21.11 podemos calcular los sumatorios necesarios para calcular el índice de Edgeworth:

$$\sum_{i=1}^{n} p_{i0}q_{i0} = 35,5$$

$$\sum_{i=1}^{n} p_{it}q_{i0} = 39,1$$

$$\sum_{i=1}^{n} p_{i0}q_{it} = 37,2$$

$$\sum_{i=1}^{n} p_{it}q_{it} = 41$$

$$IPE_0^t = \frac{39,1 + 41}{35,5 + 37,2} * 100 = 110,17\,\%$$

Como podemos observar este índice es un promedio del índice de Laspeyres y el de Paasche.

Índice de cantidades de Edgeworth

Este índice es equivalente al de precios cambiando precios por cantidades y ponderando por suma del valor en el periodo de referencia más el valor en el momento actual en el supuesto que la cantidad sea la misma que en el periodo inicial .
Ponderaciones $w_i = q_{i0}p_{i0} + q_{i0}p_{it}$

$$IQE_0^t = \frac{\sum_{i=1}^n q_{it}p_{i0} + \sum_{i=1}^n q_{it}p_{it}}{\sum_{i=1}^n q_{i0}p_{i0} + \sum_{i=1}^n q_{i0}p_{it}} * 100$$

21.6.4 Índice de Fisher

Índice de precios de Fisher

En la sección anterior se observa que dependiendo del índice que se utilice los resultados son diferentes si las cantidades consumidas no se han mantenido constantes en el tiempo. Ante este problema, el estadístico Fisher construyó un nuevo índice que proporciona un valor medio entre el índice de Lasspeyres y el índice de Paasche.

El índice de Fisher (IPF) se define como la media geométrica de los índices de Laspeyres y Paasche.

$$IPF_0^t = \sqrt{IL_0^t * IP_0^t}$$

Para el ejemplo que venimos viendo el $IF_0^t = \sqrt{110,14 * 110,22} = 110,18$

Este índice es muy poco utilizado por los economistas, ya que se necesitan los dos índices y, el índice de Paasche necesita disponer de la información actualizada tanto de los precios como de las cantidades de cada uno de los productos de la cesta, en cada uno de los periodos.

Índice de cantidades de Fisher

De forma análoga al índice de precios de Fisher, se construye el índice de cantidades de Fisher (IQF) como la media geométrica de los índices de cantidades de Laspeyres y Paasche.

$$IQF_0^t = \sqrt{IL_0^t * IP_0^t}$$

21.7 Propiedades de los principales índices complejos ponderados

Cuando se estudiaron las propiedades de los índices simples en la sección 21.3, se enumeraron como propiedades de estos índices su existencia, identidad, circular, inversión, encadenamiento, adición, multiplicación y homogeneidad.

Para los índices complejos ponderados no siempre se cumplen estas propiedades. Veamos algunos ejemplos:

Índice de Bradstreet y Dûtot cumple todas las propiedades de los números índices simples.

Índice de Suerbeck para este índice ni la propiedad de inversión ni la circular se cumplen.

$$I_0^t = \frac{\sum_{j=1}^{n} I_0^t(Y_j)}{n}$$

Índices de Laspeyres y Paasche no cumplen ni la propiedad de inversión ni la circular.

Índices de Edgeworth y Fisher la única propiedad que no se cumple es la circular.

La propiedad de proporcionalidad para los índices de Paasche, Edgeworth y Fisher se verifica pero con ciertas limitaciones. Para los índices de precios se debe suponer que al variar los precios en ciertas proporciones las cantidades no varían. Para los índices de cantidades la hipótesis es que los precios se mantienen constantes frente a las variaciones de los precios.

21.8 Índices de precios, de volumen y de valor

Como resumen, cuando la variable que se va a analizar es el precio. Los números índices más utilizados se pueden clasificar en:

- Índices de precios
  - Simples
  - Complejos
    - Sin ponderar
      - Sauerbek
      - Bradstreet-Dûtot
    - Ponderados
      - Laspeyres
      - Paasche
      - Edgeworth
      - Fisher

Los índices de precios, volumen (cantidad) o valor, asociados a las principales variables macroeconómicas son aquellos donde el valor puede definirse como la multiplicación de "precio" por "cantidad". Estos índices son clave al momento de conocer y analizar tanto la coyuntura económica y social de un país como las tendencias en el medio y largo plazo.

La estimación de esos tres índices, debe cumplir que el producto del índice de precios (IP) por el de volumen (IQ) dé el índice de valor (IV), puede basarse en diferentes enfoques metodológicos.

1. Estimación por separado los índices de precios, de cantidades y de valor, posteriormente se ajustan para que verifiquen que $IP_0^t * IQ_0^t = IV_0^t$.
2. Estimación del índice de valor y el de precios por vías diferentes y calcular el de volumen a partir de estos.

3. Estimación del índice de valor y el de volumen por vías diferentes y calcular el de precios a partir de estos.
4. Estimación integrada de los índices de precios, volumen y valor utilizando una misma vía directa o en su defecto fuentes de información articuladas de forma eficaz y eficiente. En este caso la metodología lleva necesariamente a que se cumpla la ecuación $IP_0^t * IQ_0^t = IV_0^t$.

Las recomendaciones internacionales han puesto un gran énfasis en las metodologías asociadas a la producción de índices de precios.

En particular, son muy valiosos e importantes los manuales publicados recientemente, referidos al Índice de Precios al Consumidor, los Índices de Precios al Productor y los Índices de Precios del Comercio Exterior, en cuya elaboración participaron el FMI, la OIT y otras agencias internacionales (Mundial s.f.(b),Mundial s.f.(a), Alvarez y Durán Lima 2011).

Se considera a estos manuales como un gran aporte en todo lo referente al conocimiento de la teoría de índices de precios y a las estimaciones de índices, pudiéndose encontrar en los mismos muchas recomendaciones de gran impacto a nivel práctico.

Uno de los objetivos de la creación de estos índices es apoyar las estimaciones de cuentas nacionales en lo referente a niveles y evoluciones del PIB (producto interno bruto) y el VBP (valor bruto de la producción) desagregado por ramas de actividad a precios corrientes y constantes, y las desagregaciones del PIB por el lado de la demanda y de la retribución de los factores. Además nos ayudan a analizar la evolución de los precios, donde se producen excedentes y cómo evoluciona la rentabilidad, etc.

En Economía es habitual comparar el valor de los artículos a lo largo del tiempo, pero esto es imposible salvo que la serie esté referenciada en precios constantes. Para pasar de precios corrientes a precios constantes se utilizan los números índices. Una de las principales aplicaciones de los índices de precios es eliminar el efecto que producen, las variaciones en los precios de los bienes, en las series de consumo. Todos hemos oído hablar del poder adquisitivo de los sueldos.

El término inflación hace referencia a lo que podemos adquirir con un determinado capital. Si el precio medio de la electricidad en 2020 fue de 40,37€ el megavatio-hora y el 10 de agosto del 2021 fue de 112€ el poder adquisitivo del euro se ha visto reducido al $35\,\%$. Si el precio se ha triplicado con respecto al periodo base, el índice de precios será del $300\,\%$ y el poder adquisitivo de la moneda (el euro) será la tercera parte de lo que era en el año de referencia.

Cuando se va a trabajar con series monetarias se necesita **deflacionar** la serie, es decir, dividir por un índice de precios.

Veamos a través de un ejemplo la relación entre el índice de precios y el nivel de vida.

Ejemplo.

Un determinado ayuntamiento prometió un presupuesto equivalente al de la legislatura anterior para familias vulnerables. Se supone que todas las familias vulnerables tienen idénticas preferencias.

Cesta de la compra	2016	2020
Precio de libros	20€/libro	42€/libro
Número de libros	15	8
Precio medio de los alimentos	2,5€/kilo	3,8€/kilo
Kilos de alimentos	150	300
Gasto total	675€	1.476€

Tabla 21.13: Gasto de las familias en la legislatura anterior y en la actualidad

Los gastos medios de las familias vulnerables en 2016 eran de 675€ y en 2020 es de 1.476€.

El ajuste para tener en cuenta el coste de vida es mediante el índice de valor:

$$\frac{1{,}476}{675} * 100 = 218,67\,\%$$

lo que indica que el nivel de vida ha subido un $118,67\,\%$. Por lo tanto, el presupuesto del 2020 se debe incrementar en la misma proporción para mantener el nivel de bienestar del 2016.

Veamos que ocurre si calculamos el IPC de Laspeyres.

El coste en el año 2020 de la cesta adquirida en 2016 es:

$$15 * 42 + 150 * 3,8 = 1.200€$$

precios actuales (2020) y cantidades en el año de referencia (2016).

El IPC resultante es:

$$\frac{1200}{675} * 100 = 177,78\,\%$$

un aumento del $77,78\,\%$.

Este calculo infraestima el verdadero valor del IPC porque supone que los consumidores no modifican su comportamiento respecto al consumo aunque se modifiquen los precios.

21.8.1 Índice de precios al Consumo

El índice de precios de Laspeyres más conocido es el IPC (Índice de Precios al Consumo) que se elabora en todo el mundo y que en España lo elabora el INE (Instituto Nacional de estadística) y lo publica del 12 al 14 de cada mes. Este indicador económico es utilizado para deflacionar las series monetarias españolas, para actualizar salarios, alquileres, etc. La precisión de este índice depende de su representatividad, es decir de la ponderación de cada uno de los productos que se consideran para calacularlo así como de los productos que se incluyen en la cesta.

El IPC se empezó a publicar en España en 1939 con periodo de referencia el 1936 y hasta 1976 se denominó índice de coste de vida. Hasta 2001 el IPC se calculaba como un índice de Laspeyres que utilizaba como ponderación los gastos realizados en la cesta de la compra de una familia media española en el periodo seleccionado como base (la cesta contenía 471 artículos y se recogía la información de miles de establecimientos de 177 municipios,52 capitales de provincia y 125 municipios no capitales). Este sistema revisaba, cada vez que se realizaba un cambio de base, la ponderación de los productos y la cesta de bienes según la Encuesta de Presupuestos Familiares, realizada por el INE.

En enero de 2002 entró en vigor el sistema de IPC base 2001. Es un índice más dinámico, ya que se ha convertido en un índice de Laspeyres encadenado anualmente. De esta formase se actualizan las ponderaciones más frecuentemente y se adapta mejor a la evolución del mercado. Este nuevo sistema permite la inclusión inmediata de mejoras en la metodología que ofrezcan distintos foros académicos y organismos nacionales e internacionales.

En enero de 2017 entra en vigor la base 2016. Entre las características de esta nueva base cabe destacar que incorpora la nueva clasificación europea de consumo denominada ECOICOP, lo que implica mayor desglose de la información (el número de subclases se amplía hasta 219).

En la actualidad, la cesta de la compra de las familias españolas se forma a partir de las más de 500 partidas de gasto que hay en la Encuesta Continua de Presupuestos Familiares. Se seleccionan 479 artículos (cerca de 220.000 precios) clasificados en 12 grupos. Además, se producen cambios en la cesta de la compra, se incorporan artículos como los servicios en línea de video y música, los juegos de azar o el café monodosis. Y se eliminan otros como el brandy, la videocamara o el DVD grabable.

Los 12 grupos del IPC base 2016 se subdividen en 43 subgrupos, 101 clases y 219 subclases; 57 rúbricas y 29 grupos especiales.

Grupos	2002	2003	2021
01. Alimentos y bebidas no alcohólicas	21,86	21,93	23,62
02. Bebidas alcohólicas y tabaco	3,22	3,18	3,20
03. Vestido y calzado	9,93	9,90	6,37
04. Vivienda	11,03	10,68	13,58
05. Menaje	6,36	6,41	5,94
06. Medicina	2,81	2,75	3,93
07. Transporte	15,58	15,32	12,45
08. Comunicaciones	2,57	2,73	3,73
09. Ocio y cultura	6,73	6,83	6,79
10. Enseñanza	1,74	1,67	1,66
11. Hoteles, cafés y restaurantes	11,27	11,18	11,64
12. Otros bienes y servicios	6,91	7,39	7,10

Tabla 21.14: Grupos de artículos para la elaboración del IPC y sus ponderaciones

En la Tabla 21.14 se observa como han ido variando las ponderaciones de los distintos grupos de artículos dependiendo de la incorporación al mercado de nuevos productos, de la desaparición de otros y del cambio de hábitos de los españoles.

Para poder comparar el IPC entre los países miembros de la Unión Europea desde 1997, se publica el **IPC armonizado** con una metodología común a todos los países.

Algunas limitaciones del IPC:

- Aunque las técnicas de elaboración del IPC son muy rigurosas en todos los países. Estados Unidos preocupado por que el IPC sobrevaloraba la inflación encargó, en 1996, a un grupo de expertos que elaborara un informe sobre esta cuestión (**Informe Boskin**). En este informe los expertos estimaron que el IPC americano podría estar sobrevalorando la inflación en un $1,1\,\%$ anual, en un rango de valores que iría del $0,8 al 1,6\,\%$.

  En España, el Servicio de Estudios de "la Caixa", consideró que sería muy útil disponer de un estudio similar al Informe Boskin y encargó a Fedea su elaboración. Los autores, Ruiz-Castillo, Ley e Izquierdo (Ruiz-Castillo, Ley e Izquierdo 1999), han llegado a conclusiones impactantes, puesto que estiman que el IPC podría sobrevalorar la inflación española en un $0,6\,\%$.

- Si escribimos el IPC como el promedio ponderado del IPC en cada uno de los hogares españoles, se oberva que la ponderación de los hogares es directamente proporcional al gasto del hogar. Como las pautas de consumo de los hogares más ricos son diferentes de los más pobres, no todos los hogares pesan lo mismo en el

IPC, por lo que el IPC se conoce como un **índice plutocrático**.

En Izquierdo, Ley y Ruiz-Castillo 2003 se estima que el gap plutocrático en España fue de $0,234\,\%$ en el periodo 1973-1981 y de $0,055\,\%$ entre 1991-1998. Esto indica que el precio de los productos que los ricos consumen en mayor proporción han subido más que los precios del resto de los bienes.

21.8.2 Índices de precios encadenados

Supongamos que tenemos los índices simples de la electricidad referenciado al año 2000.

Periodo	Precio	Índice (ref. 2010)
2010	45,83	100 %
2011	60,22	131,40 %
2012	59,57	129,98 %
2013	57,79	126,10 %
2014	55,05	120,12 %
2015	62,84	137,12 %
2016	48,42	105,65 %

Tabla 21.15: Índice de precios de la electricidad en España

Si se quiere cambiar el periodo base al 2016 para hacer más fácil la interpretación no se necesitan volver a calcular todos los índices, basta con dividir por el índice en base 2010, es decir, aplicar la propiedad circular.

$$I^t_{2010} = \frac{I^t_{2016} * I^{2016}_{2010}}{100}\,\%$$

$$I^t_{2016} = \frac{I^t_{2010}}{I^{2016}_{2010}} * 100\,\%$$

Periodo	Índice (ref. 2010)	Índice (ref. 2016)
2010	100 %	94,65 %
2011	131,40 %	124,37 %
2012	129,98 %	123,03 %
2013	126,10 %	119,35 %
2014	120,12 %	113,69 %
2015	137,12 %	129,78 %
2016	105,65 %	100 %

Tabla 21.16: Índice de precios de la electricidad en España

Como alternativa al cambio de base para actualizar los datos, se utiliza el sistema de índices encadenados o sistemas de bases variables en los que se usa como base el periodo inmediatamente anterior.
Para ver la diferencia entre el sistema de base fija y el sistema de encadenado se utilizará el ejemplo anterior del precio de la electricidad en España (Tabla 21.12).

Periodo	Precio	Índice (ref. 2010)	Índice encadenado
2010	45,83	100	
2011	60,22	131,40 %	131,4 %
2012	59,57	129,98 %	98,92 %
2013	57,79	126,10 %	97,01 %
2014	55,05	120,12 %	95,26 %
2015	62,84	137,12 %	114,15 %
2016	48,42	105,65 %	77,05 %

Tabla 21.17: Índice de precios de la electricidad en España encadenados

Como se puede observar en la Tabla 21.17 el precio de la electricidad en 2011 subió un $31,4\,\%$ con respecto al precio en el año 2010 y en 2015 volvió a subir un $14,15\,\%$ con respecto al 2014.

Si no tenemos un único precio, si no que tenemos una cesta de productos y estamos utilizando el índice de Laspeyres o Paasche pasar de un periodo de referencia a otro no es tan sencillo. Por lo tanto cuando estamos elaborando índices complejos no solo debemos fijar el periodo de referencia, también se debe decidir si se va a elaborar un índice encadenado o con un periodo de referencia fijo.

En el caso del IPC ya vimos que a partir del 2001 se cambió de metodología y se pasó de un periodo de referencia fijo a un periodo variable o encadenado. Esto es debido a que la estructura de gasto de las familias españolas está sometida a una constante evolución y a medida que nos alejamos del periodo de referencia se van a producir cambios bien porque los productos que se tenían en cuenta en la cesta dejan de producirse, porque aparecen nuevos productos o bien porque el consumo de ciertos productos aumenta o se restringe considerablemente.

Es importante observar que cuando se cambia de periodo de referencia no solo nos encontramos con dos periodos diferente, nos encontramos con dos cestas de productos y cantidades diferentes.
Por lo tanto si se aplica la propiedad circular o de encadenamiento de los números índices a los índices de Laspeyre o Paasche se debe de resaltar que no siempre los resultados son exactos. Si los productos de la cesta o las ponderaciones han variado de un periodo a otro el resultado será una aproximación.

21.8.3 Índices implícitos de precios

Entre los indicadores que también pueden observarse para analizar el comportamiento de los precios, se encuentra el **Índice de Precios Implícitos (**$I_0^t(PI)$), que se calcula como la diferencia entre el Producto Bruto Interno (PB) a precios constantes (en nuestro caso actual, a precios del año 1993) y el PB a precios corrientes; esto es, a los precios vigentes en el período al cual se refiere el producto. Este índice se deriva de la Contabilidad Nacional y utiliza la metodología de Paasche.

En general, los cambios que se producen en los precios puede ser debido a cambios de calidad, a la aparición de nuevos productos o a cambios producidos por la desaparición de productos ya existentes. Veamos como se pueden aislar estos cambios.

La producción bruta (PB**)** viene dada por el consumo intermedio (CI), la remuneración de asalariados (RA), los impuestos indirectos descontando los subsidios ($IImS$) y los excedentes de operaciones (EO).

$$PB = CI + RA + IImS + EO$$

Si lo pasamos a términos constantes se debe dividir por los índices de cada una de las variables.

$$\frac{PB}{I_0^t(PB)} = \frac{CI}{I_0^t(CI)} + \frac{RA}{I_0^t(RA)} + \frac{IImS}{I_0^t(IImS)} + \frac{EO}{I_0^t(EO)}$$

Por lo que **el índice de producción bruta** es la producción bruta entre su deflactor implícito. Siendo el deflactor implícito una combinación de los términos constantes de sus componentes.

$$I_0^t(PB) = \frac{PB}{\frac{CI}{I_0^t(CI)} + \frac{RA}{I_0^t(RA)} + \frac{IImS}{I_0^t(IImS)} + \frac{EO}{I_0^t(EO)}}$$

En las cuentas nacionales existen dos métodos alternativos para el cálculo de las variables macroeconómicas en valores constantes que, deben dar el mismo resultado.

1. Elaborar los índices de volumenes de tipo Laspeyres, que combinan anualmente los cambios en las cantidades y calidades de los productos al ponderarlos por sus precios en el año base. Así, el valor agregado económico se pone en función de las cantidades producidas, eliminándose de los flujos corrientes las incidencias debidas a las fluctuaciones de los precios.

2. El segundo método consiste en deflactar los valores corrientes de las series mediante los índices de precios de Paasche, con base variable. Esto requiere más información y debemos de tener en cuenta para la comparación de distintos años si las ponderaciones han variado.

El deflactor implícito de la producción o índice de precios implícitos ($I_0^t(PI)$ se define, según las cuentas nacionales como el siguiente índice de Paasche:

$$I_0^t(PI) = \frac{PB(corriente)}{PB(constantes)} = IPP_0^t = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{N} p_{i0}q_{it}} * 100$$

El origen de este cálculo es porque primero se realizan las cuantificaciones del valor de la producción, según los subgrupos preseleccionados. Este valor de la producción se sustenta en un amplio y homogéneo detalle de productos, resultantes de la actividad desarrollada por los establecimientos productores. También se añade el agregado "otros productos" generalmente integrado por un conjunto de artículos y subproductos derivados de los procesos productivos básicos.

Por lo tanto una diferencia entre el $I_0^t(PI)$ y el IPC es que el IPC es un índice de ponderaciones fijas (índice de Laspeyres), o sea que la incidencia de cada bien o servicio en la cesta total de bienes se mantiene constante a lo largo del tiempo. Por el contrario, el $I_0^t(PI)$ es un índice de ponderaciones móviles (Índice Paasche), o sea que la incidencia de cada bien o servicio se va modificando a lo largo del tiempo en función de la distinta velocidad relativa de su crecimiento productivo. Además, la composición de la cesta de bienes en ambos indicadores es diferente. El IPC comprende una cesta de productos y servicios previamente seleccionados y el $I_0^t(PI)$ incluye todos los bienes y servicios finales producidos por la economía en un período determinado. O sea que el índice de precios ímplicitos incluye un universo de bienes y servicios mucho más amplio que el IPC.

Bibliografía

Montiel Torres, AM, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson (página 112).

Ruiz-Castillo, Javier, Eduardo Ley y Mario Izquierdo (1999). *La medición de la inflación en España*. 17. "la Caixa" (página 128).

Izquierdo, Mario, Eduardo Ley y Javier Ruiz-Castillo (2003). "La brecha plutocrática en el IPC: evidencia de España". En: *IMF Staff Papers* (página 129).

Martín Guzmán, P. (2006). *Manual de estadística: descriptiva*. Thomson - Civitas (página 112).

Sánchez, José Miguel Casas, Juana Domínguez Domínguez, Carmelo García Pérez, Emilia Isabel Martos Gálvez, Luis F Rivera Galicia y Ana Isabel Zamora Sanz (2010). *Estadística para las ciencias sociales*. Editorial Universitaria Ramón Areces (página 112).

Alvarez, Mariano y José Elías Durán Lima (2011). "Manual de comercio exterior y política comercial: nociones básicas, clasificaciones e indicadores de posición y dinamismo". En: (página 125).

Mundial, Banco (s.f.[a]). "Manual del índice de precios al consumidor". En: () (página 125).

– (s.f.[b]). "Manual del índice de precios al productor". En: () (página 125).

Tema 22

Series temporales. Componentes de una serie temporal. Modelo aditivo y multiplicativo. Métodos para la determinación de la tendencia.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

AM Montiel Torres, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson

Jose María Montero Lorenzo (2007). *Estadística descriptiva*. Editorial Paraninfo

P. Martín Guzmán (2006). *Manual de estadística: descriptiva*. Thomson - Civitas

Daniel Peña (2005). *Análisis de series temporales*. Alianza

22.1 Introducción

Una serie temporal es una sucesión de observaciones realizadas, de forma secuencial, en el transcurso del tiempo. Los valores de la variable siempre van ligados a instantes de tiempo.

Toda serie temporal refleja el comportamiento de una variable en el tiempo. Normalmente las observaciones se toman a intervalos iguales de tiempo.

Las series pueden tener una periodicidad anual, semestral, trimestral, mensual, etc. Las compras anuales de una empresa, el número de casos infectados al día por un determinado virus, la cantidad de ventas diarias, etc.

El análisis de series temporales tiene dos objetivos: estudiar y modelizar el comportamiento de un fenómeno aleatorio (variable) que evoluciona a lo largo del tiempo y realizar previsiones de dicho fenómeno en el futuro. Para ello debemos suponer que las condiciones estructurales que conforman la serie objeto de estudio permanecen constantes. El enfoque clásico utilizado para alcanzar estos objetivos define a la serie temporal como la variable dependiente y el tiempo como variable soporte.

En el análisis de series podemos distinguir dos grupos de magnitudes: magnitudes stock y magnitudes flujo.

Magnitudes stock: son aquellas series que toman valores concretos en momentos concretos del tiempo.

Figura 22.1: Población en la Comunidad de Madrid

Magnitudes flujo: aquellas que representan el total acumulado de una variable desde la observación anterior.

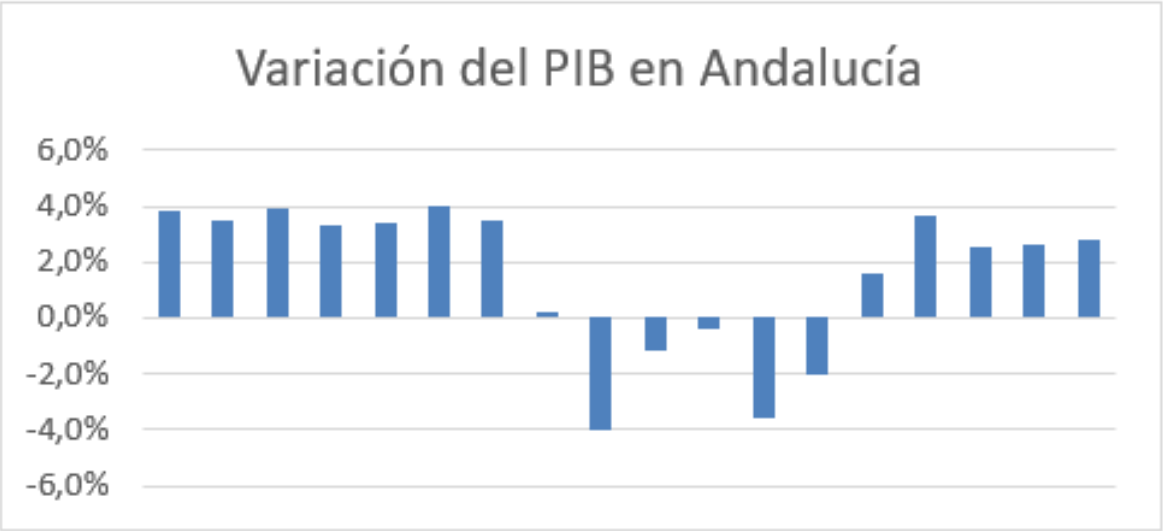


Figura 22.2: Variación del PIB en Andalucía

Mientras que el valor del flujo depende del intervalo que estemos analizando entre dos valores consecutivos, el stock no está afectado.

El análisis descriptivo de una serie es el primer paso a la hora de estudiar una serie temporal. Esta fase nos permite detectar las características más importantes de una serie, tales como su tendencia (movimiento a largo plazo), la existencia de ciclos, amplitud de las oscilaciones, presencia de valores atípicos, etc.

La forma más sencilla de comenzar el análisis de una serie temporal es mediante su representación gráfica. El gráfico que se emplea para representar las series temporales (Y_t) es el gráfico de secuencia, el tiempo se representa en el eje de abscisas (X), y la variable cuya evolución en el tiempo estudiamos en el eje de ordenadas (Y).

Figura 22.3: Secuencia de la serie del IBEX35

22.2 Componentes de una serie temporal

En el análisis descriptivo de una serie temporal, se basa en la idea de descomponer la volatilidad de la serie en varias componentes básicas. Este enfoque no siempre resulta ser el más adecuado, pero es interesante cuando en la serie se observa cierta tendencia y/o cierta periodicidad. Hay que resaltar que esta descomposición no es en general única. El objetivo por el cual se realiza la descomposición de la serie en sus componentes básicas es encontrar componentes que correspondan a una tendencia a largo plazo, un comportamiento estacional y una parte aleatoria. Consideraremos que toda serie temporal que analicemos está formada por cuatro componentes teóricas: tendencia, variaciones estacionales, variaciones cíclicas y variaciones residuales.

22.2.1 Tendencia

La determinación de la tendencia solamente se debe realizar cuando disponemos de una larga serie de observaciones.

Tendencia: es la componente no observable de la serie que nos explica el comportamiento de la variable aleatoria a largo plazo. La denotaremos por T_t y nos permitirá explicar si las medias de los valores de la serie (nivel), en una determinada unidad de tiempo, aumentan o disminuyen en el período que queremos analizar. Es de carácter determinista.

Nivel: el nivel de una serie es una medida local de tendencia central como por ejemplo

la media o mediana, de cada periodo de tiempo que consideremos.

El nivel se calcula cuando tenemos los datos en una unidad de tiempo y queremos calcular la tendencia utilizando unidades superiores.

22.2.2 Variabilidad

Además de la tendencia de una serie es interesante analizar su variabilidad mediante una medida de dispersión (varianza, recorrido intercuartílico, etc.).

Variaciones cíclicas: son variaciones que se producen con una periodicidad superior al año y frecuentemente se manifiestan como consecuencia de períodos de prosperidad y de depresión en la actividad económica, o en otras magnitudes cualquiera. Se denotan por c_{ik}.

Variaciones estacionales: son oscilaciones que se producen con una periodicidad dentro del año y que se pueden identificar repetidamente a lo largo de los años de los que disponemos datos por analizar. Por ejemplo, las ventas de helados aumentan en verano y disminuyen en invierno, Las temperaturas en Madrid tienen subidas significativas en los meses de julio y agosto y son mínimas en enero y diciembre, etc. Las denotamos por e_{ik}.

Estas variaciones se pueden medir en valores absolutos (componente estacional) o en valores relativos (índices estacionales) respecto a la media global.

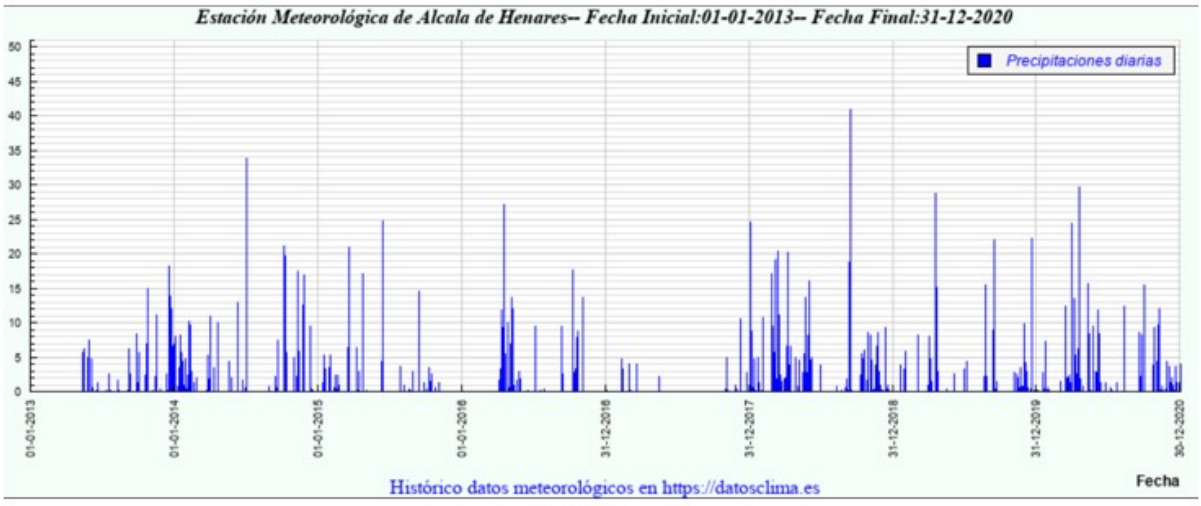


Figura 22.4: Secuencia de las precipitaciones diarias en Alcalá de Henares

Variaciones residuales: son variables aleatorias independientes con media cero y varianza constante.

Como los datos son empíricos, es de esperar que de manera natural haya en ellas pequeñas variaciones aleatorias respecto al modelo teórico que pretende analizar la serie con la información del resto de los componentes. Es necesario que no presentan ningún patrón, es decir, que sean aleatorias y de valor reducido. Estas variaciones también son llamadas ruido y debe ser aleatorio, impredecible, y con media cero. Las denotamos por r_{ik}.

22.3 Clasificaciones descriptivas de una serie temporal

1. Las series temporales se pueden clasificar en discretas y contínuas, dependiendo de que la variable que se está analizando sea discreta o continua.

2. Otra clasificación se puede realizar dependiendo de que se puedan o no predecir con exactitud los valores futuros.

   -Se llama **determinista**, si los valores futuros se pueden predecir con exactitud.

   -Si el futuro solo se puede determinar de forma parcial a partir de los valores pasados de la serie, entonces la serie temporal es **estocástica**.

3. Si la clasificación la realizamos dependiendo de la tendencia y las variaciones estacionales, nos encontramos con series estacionarias o no.

   -**Estacionarias:** una serie es estacionaria cuando es estable a lo largo del tiempo, es decir, cuando la media y varianza son constantes en el tiempo. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo.

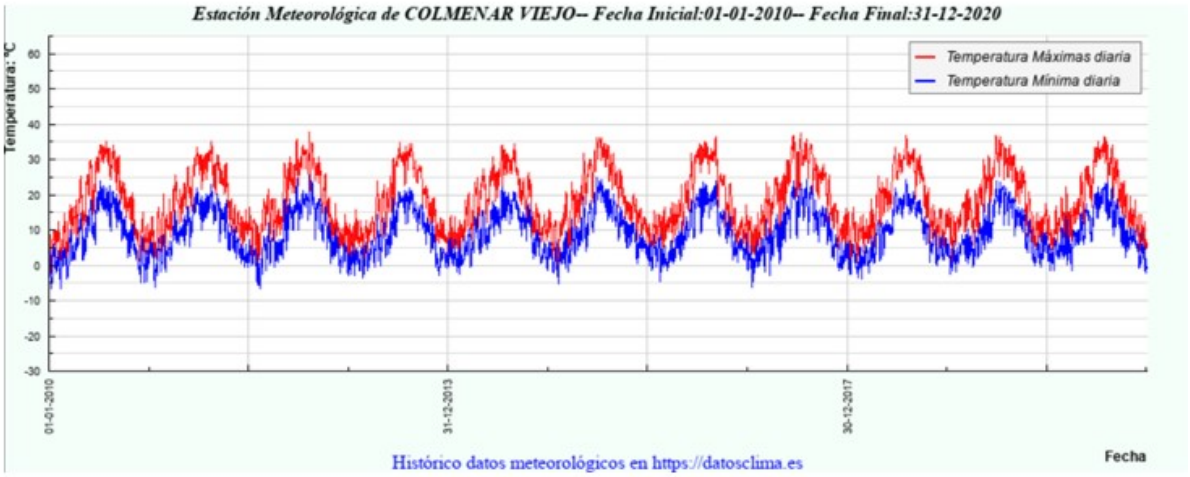


Figura 22.5: Serie Temperaturas medias en el municipio de Colmenar Viejo durante diez años

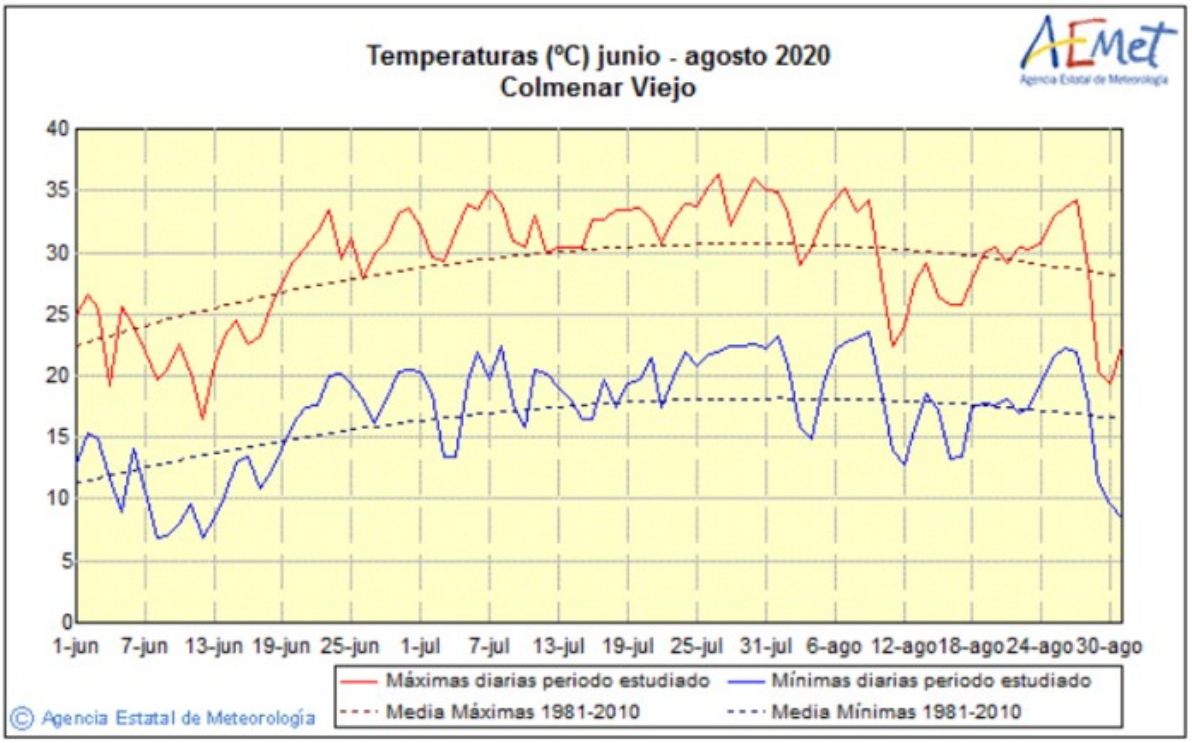


Figura 22.6: Serie Temperaturas medias en el municipio de Colmenar Viejo durante un trimestre

Como se observa en la Figura 22.5 la temperatura media es constante si nos fijamos en unidades de tiempo como el año, la serie es estable alrededor de un valor central. Sin embargo, si tomamos unidades de tiempo más pequeñas, como los días, esa media va modificándose y no se observa la estacionalidad de la serie. Lo mismo ocurre con la varianza.

-**No estacionarias:** son series en las cuales la tendencia y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante.

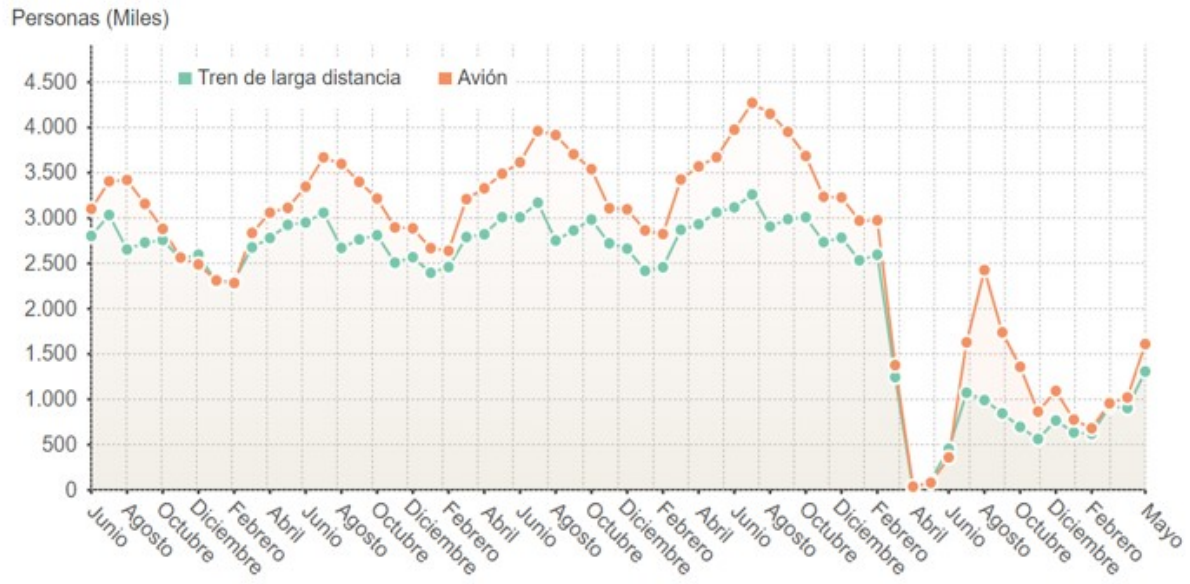


Figura 22.7: Viajeros en avión y en tren de larga distancia desde junio del 2016 hasta mayo del 2021. Fuente: www.epdata.es

En este gráfico se observa como la media anual es creciente hasta 2019, y en 2020 y 2021 cambia la tendencia y la variabilidad.

4. Dependiendo de la estructura con la que se unen los distintos componentes de la serie, se pueden clasificar en series con modelo aditivo y series con modelo multiplicativo.

-**Modelo aditivo:** si la variabilidad de una serie no depende del nivel significa que los componentes de la serie se combinan de forma aditiva, es decir, el incremento debido a la estacionalidad siempre es el mismo, aunque exista tendencia creciente o decreciente. Esquema aditivo:

$$Y_{i,t} = T_{i,t} + c_{i,t} + e_{i,t} + r_{i,t}$$

-**Modelo multiplicativo:** si la variabilidad y el nivel dependen entre sí los elementos de la serie se combinan de forma multiplicativa. Esto quiere decir que el incremento debido a la estacionalidad aumenta o disminuye conforme la tendencia crece o decrece. Esquema multiplicativo:

$$Y_{i,t} = T_{i,t} * c_{i,t} * e_{i,t} * r_{i,t}$$

Se puede observar que es posible pasar de un modelo multiplicativo a un modelo aditivo sin más que aplicar logaritmos neperianos.

$$Ln(Y_t) = Ln(T_t) + Ln(c_t) + Ln(e_t) + Ln(r_t)$$

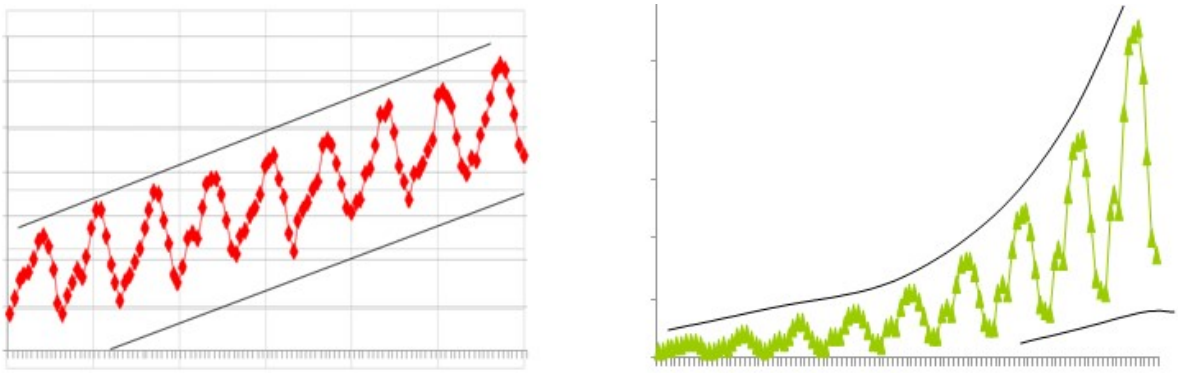


Figura 22.8: Representación gráfica de dos series con distinto esquema de combinación entre sus componentes de tendencia y estacionalidad

En el modelo multiplicativo la tendencia se expresa en el mismo tipo de unidad que las observaciones, y el resto de las componentes en tanto por uno. Sin embargo, en el modelo aditivo todas las componentes se expresan en las mismas unidades que las observaciones.

Para estudiar el esquema de la serie temporal debemos analizar si existe dependencia entre variabilidad y el nivel. Una forma de observar esta dependencia es representando el gráfico de dispersión en el que se representa el logaritmo neperiano de la mediana (u otra medida de tendencia central) frente al logaritmo neperiano del recorrido intercuartílico (también se puede utilizar la varianza) de cada uno de los periodos considerados en la serie.

Otra forma de saber si una serie es aditiva o multiplicativa es analizar la serie de los cocientes C_i y de las diferencias D_i:

$$C_i = \frac{Y_{i,k+1}}{Y_{i,k}},$$

donde C_i es el cociente entre dos datos del mismo periodo i correspondiente a dos años consecutivos k y $k+1$.

$$D_i = Y_{i,k+1} - Y_{i,k},$$

siendo D_i la diferencia entre dos datos del mismo periodo i correspondiente a dos años consecutivos k y $k+1$ y el subíndice i hace referencia al periodo i-ésimo del año k.

Para cada una de estas variables se calcula el coeficiente de variación, de forma que:

- Si $CV(C) > CV(D)$, la serie presenta un esquema aditivo.

- En caso contrario, si $CV(D) > CV(C)$, el esquema será multiplicativo.

Si nos fijamos en la serie de compraventas de viviendas en España:

Año	Primer Trimestre	Segundo Trimestre	Tercer Trimestre	Cuarto Trimestre
2007	230.023	202.585	184.326	158.366
2008	162.267	150.981	128.773	110.059
2009	106.523	97.788	109.084	99.998
2008 Cocientes C_i	0,705	0,745	0,699	0,695
2009 Cocientes C_i	0,656	0,648	0,847	0,909
2008 Diferencias D_i	-67.756	-51.604	-55.553	-48.307
2009 Diferencias D_i	-55.744	-53.193	-19.689	-10.061

Tabla 22.1: Serie de las compraventas de viviendas en España

Primero calculamos la variable C_i y la variable D_i y luego para estas variables calculamos su coeficiente de variación.

$$CV(C) = \frac{S_C}{\bar{C}} = 0,1176$$

$$CV(D) = \frac{S_D}{\bar{D}} = 0,4081$$

como

$$CV(D) > CV(C),$$

la serie presenta un esquema multiplicativo.

22.4 Cálculo de la tendencia

Calcular la tendencia de una serie es despojar a la serie del resto de sus componentes. Para calcular la tendencia podemos utilizar distintos métodos: gráfico o visual, de medias móviles o de ajuste analítico.

22.4.1 Método gráfico

Una vez calculada la media en la unidad de tiempo definida, representamos la serie de las medias para determinar si esta serie es o no estable. En el caso de la serie del IBEX

vemos que no lo es, la media del valor del IBEX contabilizado cada día (periodo elegido para dividir el calendario) va aumentando progresivamente, de manera que presenta tendencia creciente.

Otra forma de estudiar el nivel de una serie es realizando un box-plot[1] (diagrama de cajas) de cada uno de los periodos de tiempo considerados.

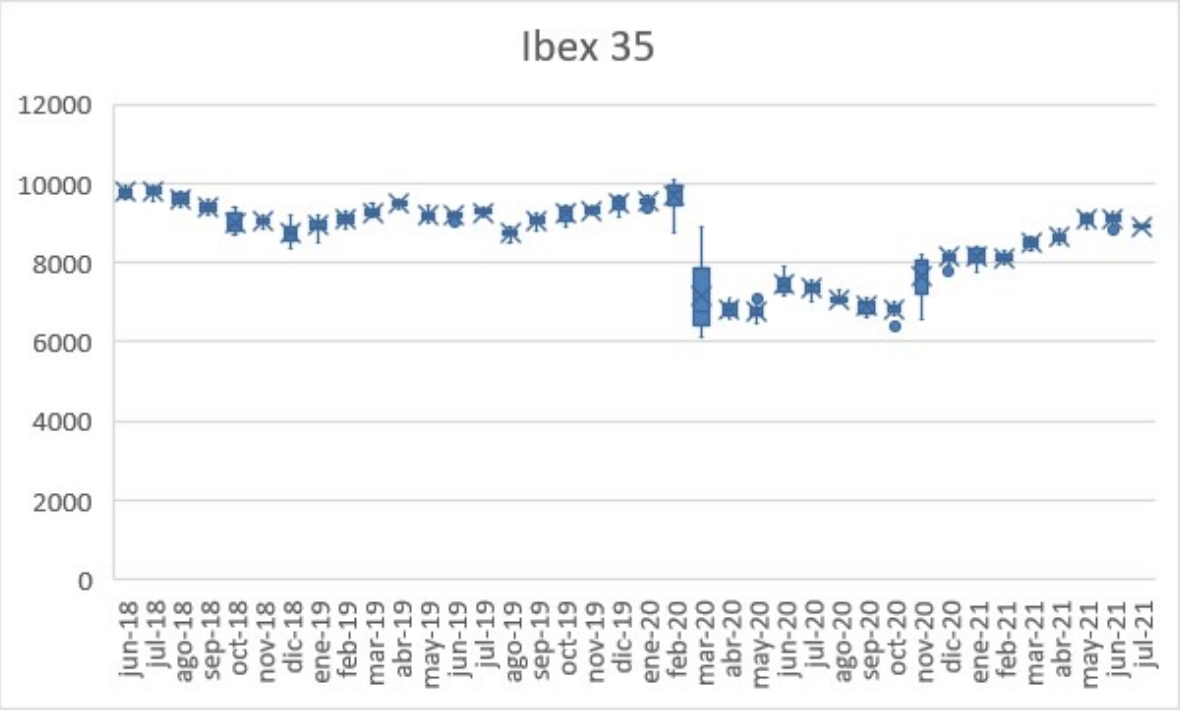


Figura 22.9: Diagramas de cajas por meses del IBEX 35

Al igual que en el caso anterior, donde representábamos el IBEX, llegamos a la conclusión de que el nivel es inestable y presenta un cambio de tendencia debido a la pandemia.

[1]El diagrama de caja (box-plot) de una variable consiste en un rectángulo (caja), cuyo extremo inferior es el primer cuartil de la variable, y el superior el tercero. Este rectángulo está dividido por un segmento que representa el segundo cuartil (mediana). Las dos líneas (bigotes) que parten de los extremos del rectángulo intentan alcanzar los valores mínimo y máximo, pero su longitud no puede superar 1,5 veces el rango intercuartílico. Si existe algún valor que queda fuera de los bigotes se representa a parte, con un asterisco.

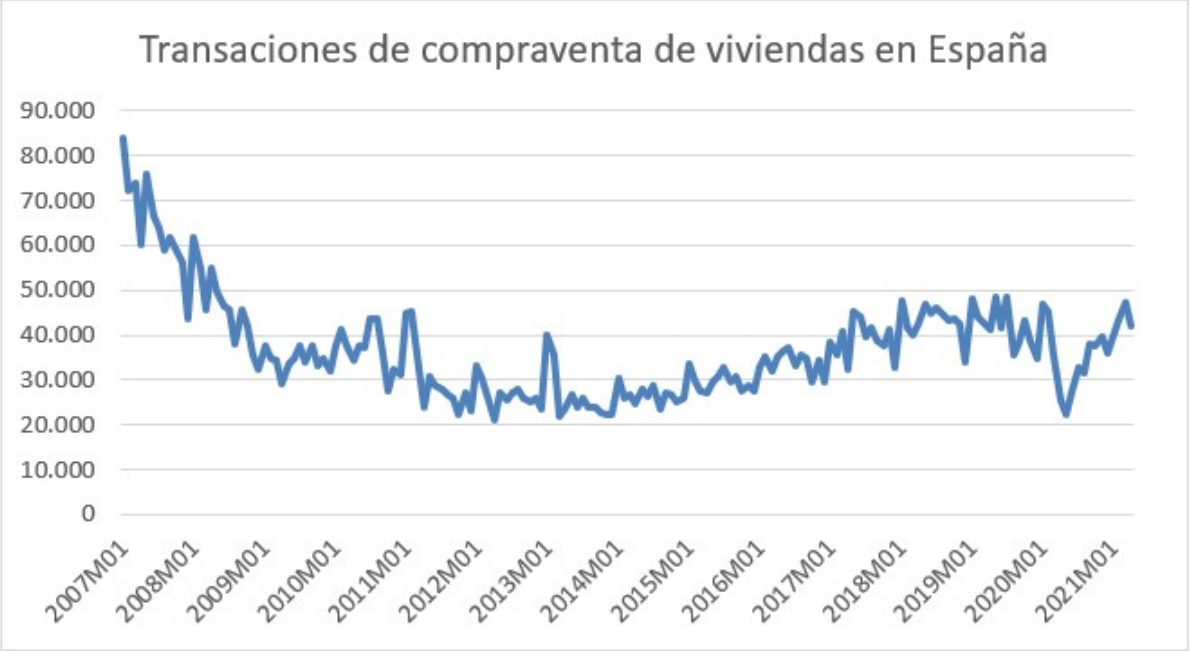


Figura 22.10: Transaciones de compraventa de viviendas en España

22.4.2 Método de medias móviles

Este método es local y no exige la suposición de una forma funcional para la tendencia. Este método además de para obtener la tendencia se utiliza para transformar la serie temporal en otra equivalente pero más suavizada. Generalmente entenderemos por suavizamiento de la serie la obtención de unos valores transformados con menos fluctuación.

Dado un conjunto de observaciones correspondientes a una serie, las medias móviles tratan de promediar de forma móvil los datos.

Medias móviles de amplitud q: un proceso de medias móviles de orden q es un proceso por el cual la tendencia de Y_t se calcula ajustando la serie obtenida como la media aritmética de q valores consecutivos de Y_t.

1. Para q impar:
   Se utiliza sobre todo cuando los datos vienen dados en semanas $(q = 7)$, trimestres $(q = 3)$, etc. La nueva serie estaría formada por:

$$Y_{(\frac{q-1}{2})+k} = \frac{Y_{1+k} + Y_{2+k} + ... + Y_{q+k}}{q}$$

Máximos del IBEX	Medias móviles $q = 5$
9741,3	
9669,2	
9723,9	9714,42
9699,8	9726,06
9737,9	9761,02
9799,5	9791,6
9844	9808,16
9876,8	9832,48
9782,6	9869,76
9859,5	9895,74
9985,9	9905,16
9973,9	9940,96
9923,9	9952,1
9961,6	9909,36
9915,2	9902,7
9772,2	9876,68
9940,6	9869,24
9793,8	
9924,4	

Tabla 22.2: Ajuste mediante medias móviles de orden $q = 5$

Figura 22.11: Cierres del IBEX y serie suavizada mediante medias móviles de orden 5

2. Para q par:
   Se utiliza cuando los datos vienen dados en meses ($q = 12$), cuatrimestres ($q = 4$), etc.

   Cuando q es par el punto medio no coincide con ninguno de los periodos para los que tenemos datos y debemos de realizarlo en dos fases, primero se realizan las medias móviles como en el caso impar y después se calcula una nueva media móvil de orden 2.

Fecha	Compraventa de Viviendas Trimestrales	Medias Móviles orden 4	Medias Móviles orden 2
2007 1T	230.023		
2007 2T	202.585		
		193.825	
2007 3T	184.326		185.356
		176.886	
2007 4T	158.366		170.436
		163.985	
2008 1T	162.267		157.041
		150.097	
2008 2T	150.981		144.058
		138.020	
2008 3T	128.773		131.052
		124.084	
2008 4T	110.059		117.435
		110.786	
2009 1T	106.523		108.325
		105.864	
2009 2T	97.788		104.606
		103.348	
2009 3T	109.084		
2009 4T	99.998		

Tabla 22.3: Ajuste mediante medias móviles de orden $q = 4$

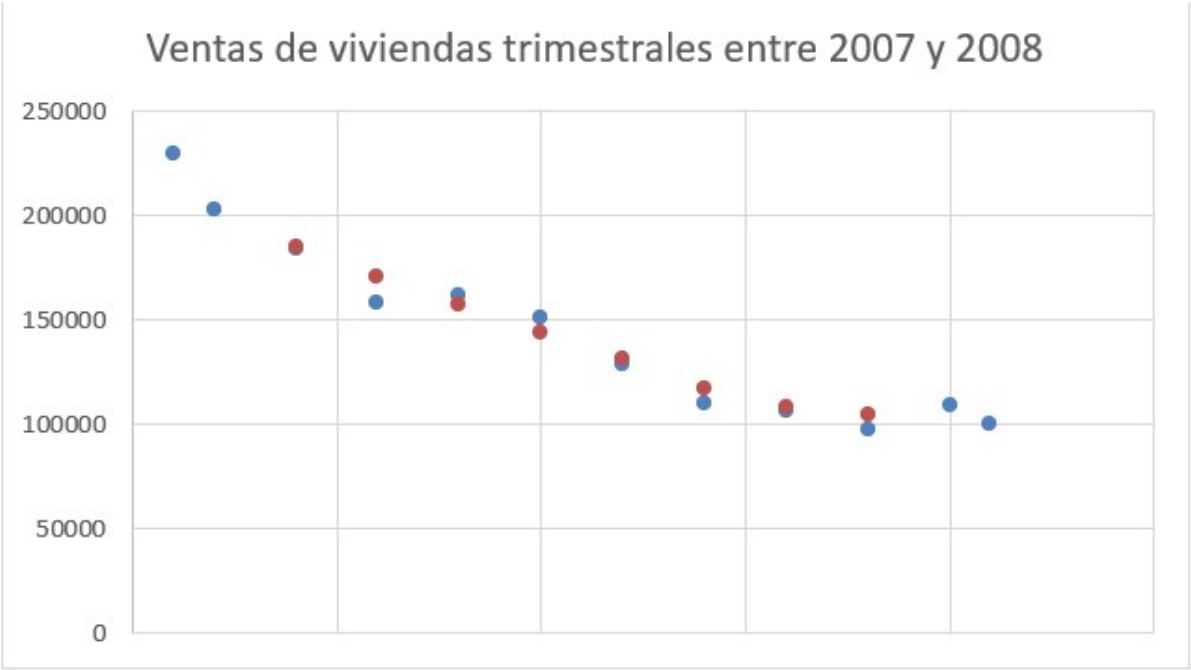


Figura 22.12: Compraventas de viviendas trimestrales en España entre 2007 y 2008

Hay que analizar cómo están medidos los datos ya que, si el orden de las medias móviles es correcto, el método puede ser útil para descubrir la tendencia. Si se tiene en cuenta que las variaciones cíclicas se repiten cada cierto período, y que la estacionalidad también se repite cada año, el cálculo de las medias móviles, en el supuesto de que el "período"de los ciclos sea un número entero de años, compensará las variaciones cíclicas y estacionales promediando las positivas y negativas; e igualmente promediará las variaciones erráticas que hay que suponer que tienen una media de cero.

22.4.3 Método de ajuste analítico

Se tratará de obtener una función que sea capaz de explicar con una buena aproximación a largo tiempo el comportamiento de la serie en función de la variable tiempo. Primero será necesario escoger el tipo de función (lineal, polinómica, exponencial, etc.) y luego habrá que determinar los parámetros de ajustes. Para escoger el tipo de función, la decisión puede basarse en el método gráfico anteriormente descrito. Y en cuanto a la determinación de la función concreta de ajusta lo más habitual será utilizar el método de mínimos cuadrados. En este tema solamente se desarrollará el ajuste lineal, que viene expresado por la siguiente ecuación:

$$Y_t = a + bt$$

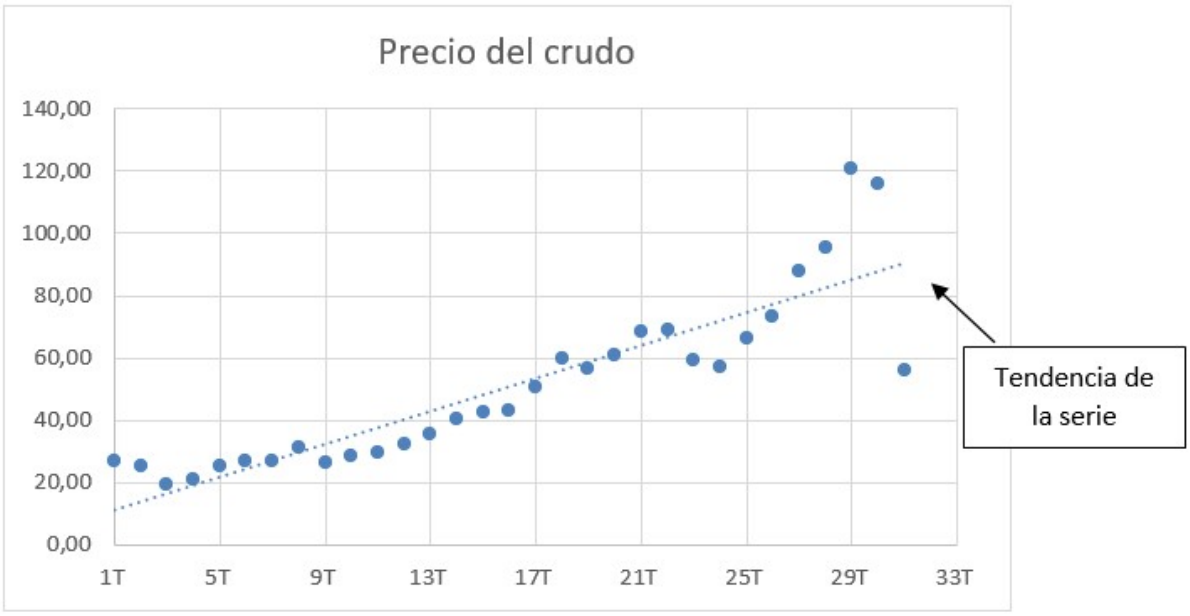


Figura 22.13: Precio medio del crudo trimestalmente

Para estimar la tendencia el primer paso que debemos realizar es transformar la variable Tiempo en las unidades que nos interesan, años, meses, etc.

Si los datos están en una unidad menor, se deben agrupar para evitar las variaciones estacionales.

$$\hat{y}_t - \bar{y} = a + \frac{cov(Y,T)}{S_T}t$$

donde:
T es la variable Tiempo,
$cov(Y,T)$ es la covarianza entre la serie temporal y la variable Tiempo y,
S_T es la desviación típica de la variable T.

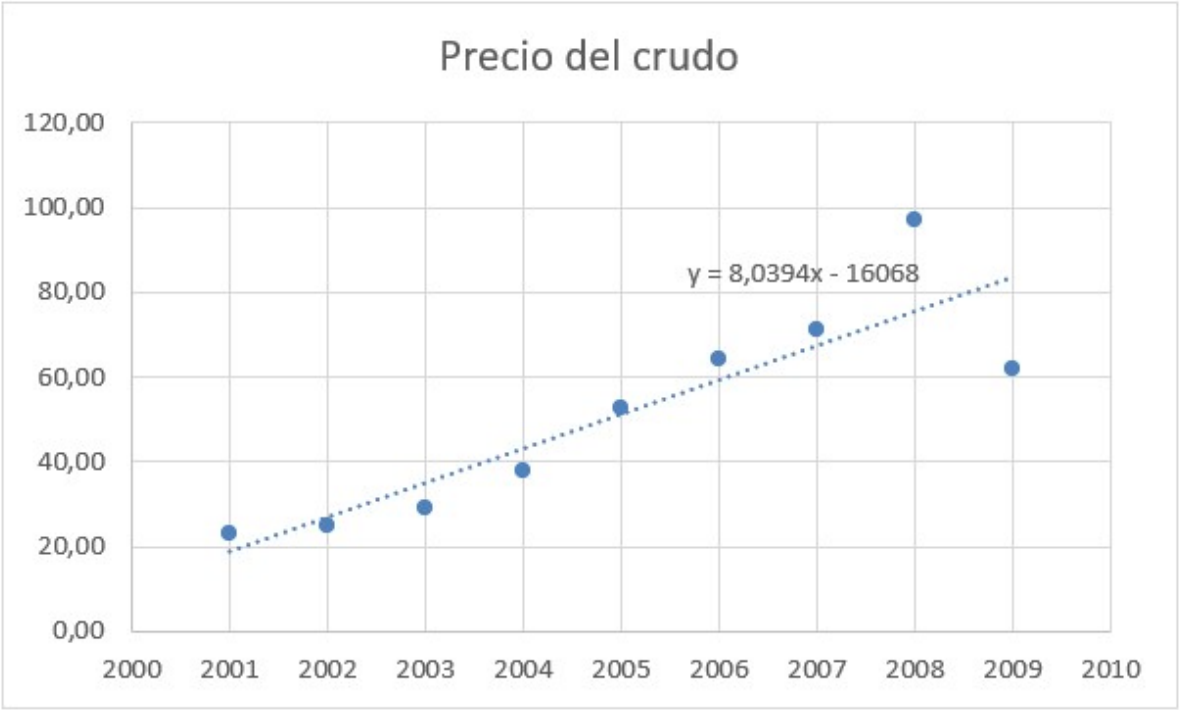


Figura 22.14: Precio medio del crudo anualmente

La pendiente de la recta indica lo que el precio del crudo sube al año. Si queremos saber lo que sube al trimestre, dividiendo este dato entre cuatro y obtenemos la diferencia entre dos valores trimestrales. Por lo tanto, lo único que hay que hacer es determinar el primero de estos valores.

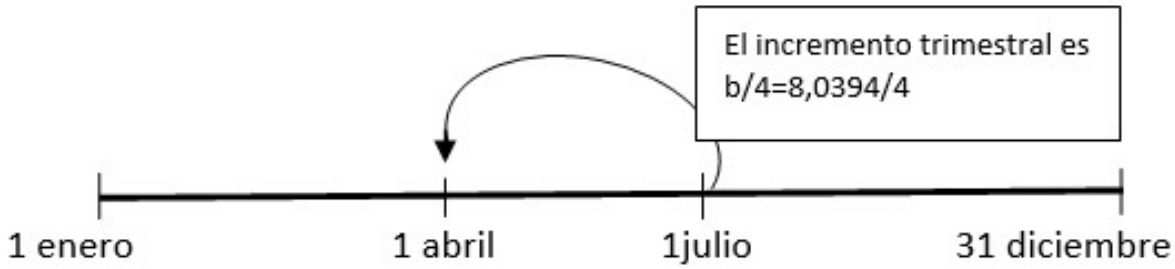


Figura 22.15: Incrementos inferiores a la unidad temporal

22.4.4 Suavizado exponencial

El suavizamiento exponencial emplea un promedio ponderado de los valores de la serie temporal pasados para predecir sus valores en periodos futuros.

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \;\; con \;\; \alpha \in (0,1)$$

En este caso solo se tiene en cuenta el valor de la serie en el instante anterior y su estimación. En la práctica comenzamos haciendo que $\hat{Y}_{t+1}$, el primer valor estimado de la serie, sea igual a Y_1, que es el primer valor real de la serie.

$$\hat{Y_3} = \alpha Y_2 + (1-\alpha)\hat{Y_2} = \alpha Y_2 + (1-\alpha)Y_1$$

$$\hat{Y_4} = \alpha Y_3 + (1-\alpha)\hat{Y_3} = \alpha Y_3 + \alpha(1-\alpha)Y_2 + (1-\alpha)^2 Y_1$$

Para el caso de compraventa de viviendas en España con $\alpha = 0,4$, obtenemos los siguientes resultados:

Fecha	Compraventa de Viviendas	Estimaciones (Serie 2)
2007 1T	230.023	230.023
2007 2T	202.585	230.023
2007 3T	184.326	219048
2007 4T	158.366	205159
2008 1T	162.267	186442
2008 2T	150.981	176772
2008 3T	128.773	166456
2008 4T	110.059	151383
2009 1T	106.523	134853
2009 2T	97.788	123521
2009 3T	109.084	113228
2009 4T	99.998	111570

Tabla 22.4: Ajuste mediante suavizado exponencial

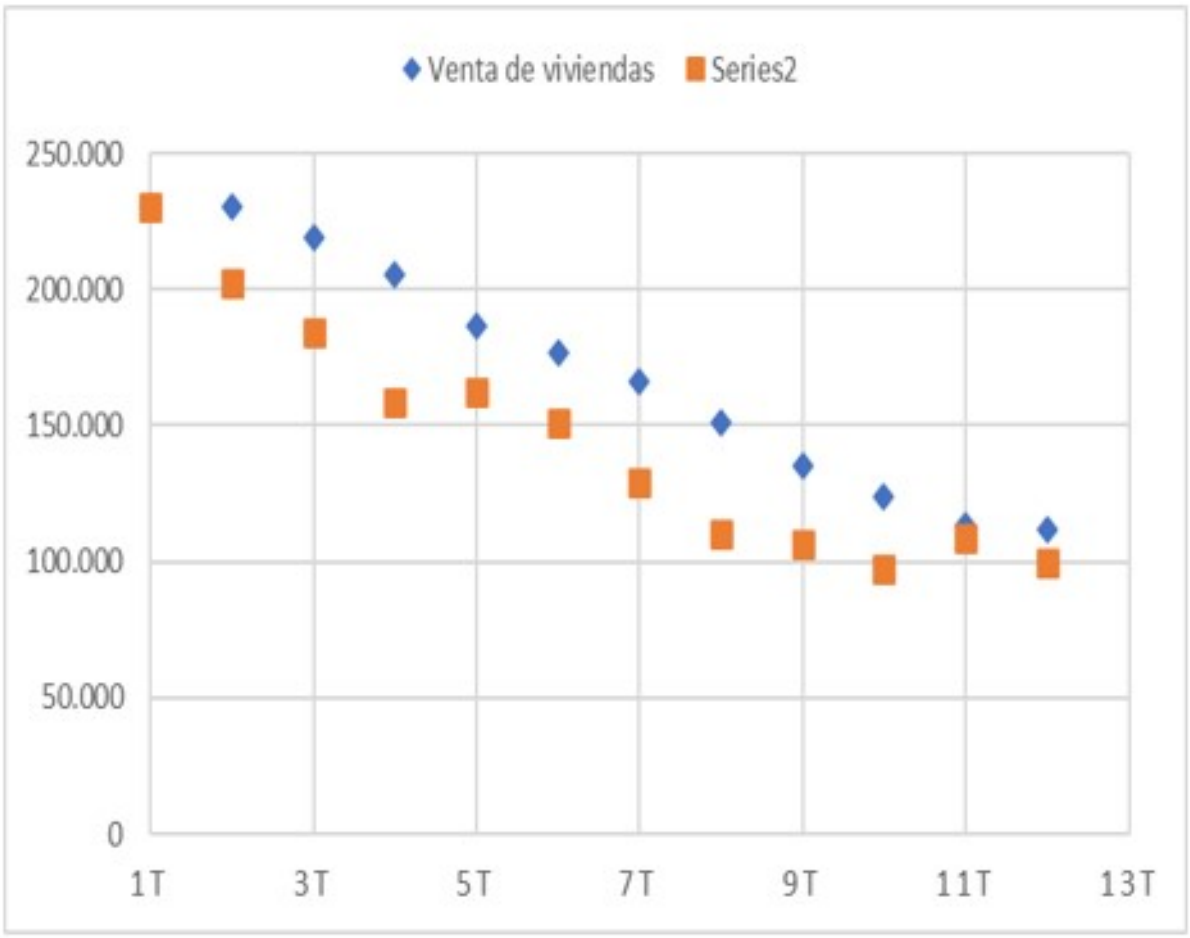


Figura 22.16: Tendencia de la serie Compraventas de viviendas mediante suavizado exponencial

Para ampliar el análisis del precio de la vivienda consultar: Puerto y Paz 2004

Bibliografía

Montiel Torres, AM, FJ Barón López y F Rius Díaz (1997). *Elementos básicos de estadística económica y empresarial*. Editorial Thomson (página 133).

Puerto, Justo y María Paz (2004). "Análisis descriptivo de series temporales aplicadas al precio medio de la vivienda en España". En: *Management Mathematics for European Schools* (página 152).

Peña, Daniel (2005). *Análisis de series temporales*. Alianza (página 133).

Martín Guzmán, P. (2006). *Manual de estadística: descriptiva*. Thomson - Civitas (página 133).

Montero Lorenzo, Jose María (2007). *Estadística descriptiva*. Editorial Paraninfo (página 133).

Tema 23

El análisis de las series temporales. Métodos elementales para la determinación de las variaciones estacionales y los movimientos cíclicos.

Este tema está elaborado como una adaptación de la siguiente bibliografía:

Daniel Peña (2005). *Análisis de series temporales*. Alianza

María Pilar González Casimiro (2009). "Análisis de series temporales: Modelos ARIMA". En

Teresa Villagarcía (2018). "Series temporales". En: *Obtenido de http://www.est. uc3m. es/esp/nueva_docencia/leganes/ing_industrial/estadistica_industrial/doc_grupo1/archivos/Apuntes % 20de % 20series. pdf*

Esta documentación es orientativa y no es exclusiva ni única para el correcto desarrollo de este tema. Tampoco vincula al órgano convocante ni al Tribunal actuante.

Aviso: El INE se reserva las acciones legales que pudieran corresponder por la venta de esta información.

23.1 Introducción

Las series pueden ser deterministas o estocásticas. Pero en la práctica, la mayoría de las series son estocásticas, es decir, los futuros valores sólo se pueden determinar parcialmente por sus valores pasados, y estas predicciones deben ser reemplazadas por la idea de que los futuros valores tienen una distribución de probabilidad que está condicionada al conocimiento de los valores pasados.

En el tema anterior se detallaron las componentes de una serie temporal (tendencia, variación estacional, variación cíclica y variación residual) y se calculó su tendencia.

En este tema se van a analizar el resto de los componentes (variación estacional y variación residual). Tanto la tendencia como la parte estacional son componentes deterministas, por lo tanto es importante aislarlas para analizar la parte aleatoria o ruido de la serie.

La componente estacional de la serie temporal es muy importante porque nos da información sobre el comportamiento de la variable a corto plazo. Y si se elimina de la serie, se puede detectar su comportamiento a largo plazo. Para identificar esta componente lo

primero que se debe hacer es eliminar la tendencia de la serie. En este tema vamos a suponer que en todas las series se ha eliminado la tendencia previamente.

Una vez identificada la tendencia y la variación estacional nos queda la parte aleatoria de la serie que se puede abordar tanto desde el punto de vista descriptivo como mediante el enfoque de Box-Jenkins.

23.2 Índice de variación estacional

La variación estacional indica el incremento o disminución de la serie respecto a la tendencia en periodos de tiempo menores a un año. Son movimientos periódicos que se repiten en intervalos cortos de tiempo con duraciones casi constantes.

Si el modelo de la serie es aditivo, la componente estacional indica la cantidad en que se ha superado o no se ha alcanzado el valor de la tendencia. Sin embargo, si el modelo es multiplicativo, la componente estacional se valora a través de un índice de variación estacional, que viene expresado en porcentaje y que significa la fluctuación del valor de la serie respecto a la tendencia.

23.2.1 Estimación determinista de la variación estacional

Para estimar la estacionalidad de la serie lo primero que debemos hacer es representarla una vez eliminada la tendencia, y así identificar el periodo de la componente estacional que puede ser semanal, mensual, trimestral...

En este apartado vamos a desarrollar el Método de la razón o de la diferencia, dependiendo que la serie temporal sigua un modelo multiplicativo o aditivo.

Modelo multiplicativo

Si partimos de un modelo multiplicativo

$$Y_{i,t} = T_{i,t} * c_{i,t} * e_{i,t} * r_{i,t}$$

Para aislar la componente estacional construimos el índice de variación estacional (IVE_i).

Previamente a la construcción del Índice de variación estacional, debemos calcular el Coeficiente de variación estacional (CVE).

Coeficiente de variación estacional: es el cociente entre el valor de la serie temporal en un determinado periodo y la tendencia. Este coeficiente nos va a servir para comprobar si en cada periodo la serie se comporta mejor o peor que la tendencia.

Para calcular este coeficiente primero se calcula el cociente

$$\frac{Y_{i,t}}{T_{i,t} * c_{i,t}} = e_{i,t} * r_{i,t}$$

y posteriormente para eliminar la parte aleatoria o residual $r_{i,t}$ se calcula la media de los cocientes para cada uno de los periodos (i).

Índice de variación estacional: recogen el incremento o la disminución porcentual que el componente estacional produce en cada estación anual.

Modelo aditivo

Por el contrario, si el modelo es aditivo

$$Y_{i,t} = T_{i,t} + c_{i,t} + e_{i,t} + r_{i,t}$$

para aislar la componente estacional lo que calculamos son las diferencias

$$Y_{i,t} - T_{i,t} - c_{i,t} = e_{i,t} + r_{i,t}$$

igual que en el método multiplicativo, posteriormente se eliminará la componente residual calculando la media aritmética en cada periodo.

Ejemplo 1. En el ejemplo de las transaciones de compraventa de viviendas en España que se analizó en el Tema 22 se vio que el modelo que rige la serie es multiplicativo (Tabla 22.1), por lo tanto debemos calcular la razón

$$\frac{Y_{i,t}}{T_{i,t} * c_{i,t}}.$$

Como los datos observados están de forma trimestral, primero se calcularon las medias móviles de orden 4 para eliminar la tendencia (Tabla 22.3).

Año	Primer Trimestre	Segundo Trimestre	Tercer Trimestre	Cuarto Trimestre
2007	230.023	202.585	184.326	158.366
2008	162.267	150.981	128.773	110.059
2009	106.523	97.788	109.084	99.998
2010	116.483	109.139	123.241	90.728
2011	122.933	83.440	80.736	72.715
2012	88.884	73.857	80.964	74.829

Tabla 23.1: Valores de la serie de transaciones de compraventa de viviendas en España

Año	Primer Trimestre	Segundo Trimestre	Tercer Trimestre	Cuarto Trimestre
2007			185.356	170.436
2008	157.041	144.058	131.052	110.059
2009	108.325	104.606	104.593	107.257
2010	110.446	111.057	110.704	108.298
2011	99.772	92.208	85.700	80.246
2012	79.077	79.369		

Tabla 23.2: Valores de la tendencia mediante ajuste de medias móviles de orden $q = 4$

Año	Primer Trimestre	Segundo Trimestre	Tercer Trimestre	Cuarto Trimestre
2007			0,994	0,929
2008	1,033	1,048	0,983	0,696
2009	0,983	0,935	1,043	0,932
2010	1,055	0,983	1,113	0,838
2011	1,232	0,905	0,942	0,906
2012	1,124	0,931		
Medias (CVE)	1,085	0,960	1,015	0,860

Tabla 23.3: Coeficientes de variación estacional (en cada casilla podemos observar el cociente entre el valor de la serie temporal y la tendencia, calculada ésta por el método de las medias móviles de orden $q = 4$)

La razón entre el valor de la serie temporal y la tendencia, calculada en la Tabla 23.3 nos facilita el producto de la componente estacional por la componente residual.

Para estimar la componente estacional se calcula la media de la razón anterior para cada uno de los trimestres y así la aleatoriedad de la componente residual desaparece. Por ejemplo, la componente estacional para el primer trimestre es la media de los valores: 1,033; 0,983; 1,055; 1,232 y 1,124. El valor de la componente estacional para el primer trimestre es 1,085 (Tabla 23.3).

De esta forma se observa que durante el primer y tercer trimestre las ventas están por encima de la tendencia ($e_1 = 1,085$ y $e_3 = 1,015$), y en el segundo y cuarto trimestre las ventas se sitúan por debajo de la tendencia ($e_2 = 0,96$ y $e_4 = 0,86$).

Por último, se calculan los índices de variación estacional como el porcentaje que supone cada coeficiente de variación estacional frente a la variación estacional media.

$$\bar{e} = \frac{\sum_1^4 e_i}{4}$$

$$IVE(i) = \frac{e_i}{\bar{e}} * 100$$

$IVE(1) = 110,737\,\%$, $IVE(2) = 97,956\,\%$, $IVE(3) = 103,552\,\%$ y $IVE(4) = 87,755\,\%$

Se han calculado las variaciones estacionales para una serie temporal con modelo multiplicativo. Si la serie temporal sigue un modelo aditivo, lo que se calcula no es la razón entre los valores de la serie y la tendencia, si no las diferencias.

Ejemplo 2. Con los datos de una serie temporal cuyo modelo es aditivo se va a aislar la componente estacional.

Supongamos que tenemos la siguiente serie:

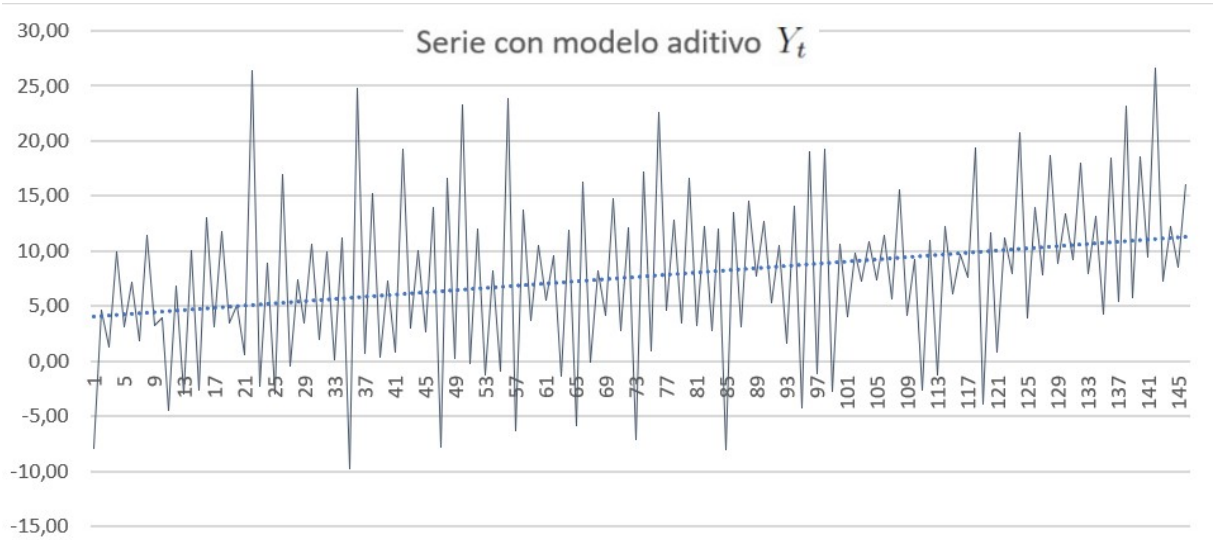


Figura 23.1: Representación gráfica de una serie temporal con modelo aditivo

Lo primero que se debe hacer es restar la tendencia. Recordemos que en los modelos aditivos la componente estacional tiene las mismas unidades que la serie.

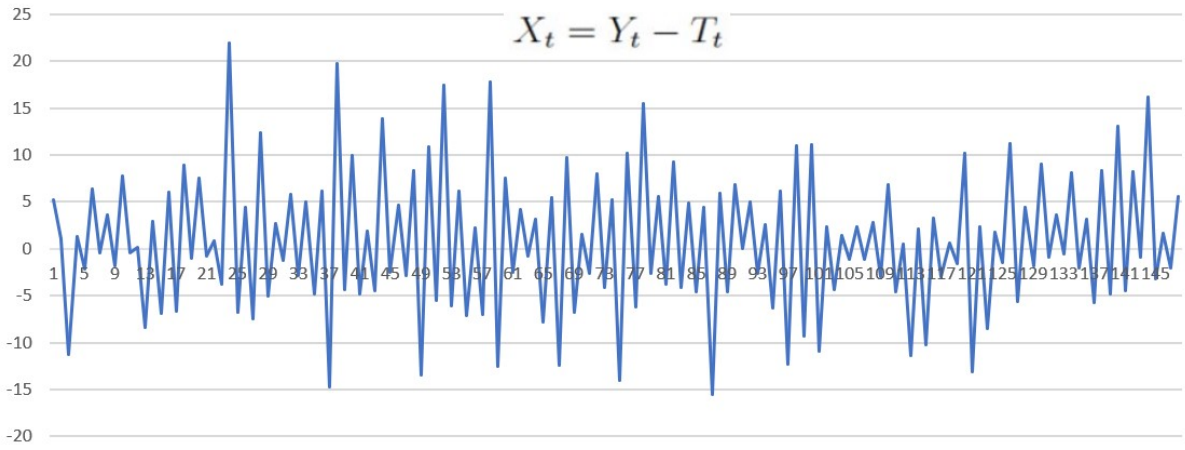


Figura 23.2: Representación gráfica de una serie temporal desestacionalizada

Bajo la hipótesis de estacionalidad estable, la diferencia entre la componente de estacionalidad y el valor de la serie X_t (donde $X_t = Y_t - T_t$) es debido a la componente residual o aleatoria.

Por ello, para estimar la componente estacional se calcula la media aritmética para cada trimestre del año. De esta forma, se elimina la componente aleatoria.

Año	Primer Trimestre	Segundo Trimestre	Tercer Trimestre	Cuarto Trimestre
2012	-2,111	6,381	-0,441	3,584
2013	-1,784	7,706	-0,452	0,159
2014	-8,344	2,896	-6,893	6,023
2015	-6,660	8,938	-0,985	7,574
2016	-0,765	0,779	-3,759	21,937
2017	-6,712	4,393	-7,487	12,383
2018	-5,093	2,717	-1,217	5,814
2019	-2,860	4,994	-4,775	6,186
2020	-14,762	19,675	-4,381	9,980
Medias	-5,455	6,498	-3,377	8,182

Tabla 23.4: Coeficientes de variación estacional (en cada casilla podemos observar la diferencia entre el valor de la serie temporal y la tendencia, calculada ésta por el método de las medias móviles de orden $q = 12$)

A partir de las medias calculadas para cada trimestre en la Tabla 23.4 se calcula el índice de variación estacional, es decir, la influencia del trimestre en el resultado de la serie.

Como nuestro modelo es aditivo, $IVE(i)$ se calcula como

$$e_i - \bar{e}, \text{ con } \bar{e} = \frac{\sum_1^4 e_i}{4}.$$

Y en nuestro ejemplo, se obtienen los siguientes resultados:

$IVE(1) = -6,917$, $IVE(2) = 5,036$, $IVE(3) = -4,839$ y $IVE(4) = 6,720$

Recordemos que ahora IVE está medido en las mismas unidades que los datos de la serie.

Por otra parte, se observa que durante el primer y tercer trimestre la serie está por debajo de la tendencia ($IVE(1) = -6,917$ y $IVE(3) = -4,839$), y en el segundo y cuarto trimestre la serie se sitúa por encima de la tendencia ($IVE(2) = 5,036$ y $IVE(4) = 6,720$).

23.3 Métodos elementales para la determinación de los movimientos cíclicos

Las variaciones cíclicas son movimientos irregulares que se producen con una periodicidad superior al año y frecuentemente se manifiestan como consecuencia de períodos de prosperidad y de depresión en la actividad económica, o en otras magnitudes cualquiera. Las denotamos por $c_{i,k}$.

Para aislar estas variaciones en un modelo multiplicativo procedemos de una forma similar al que realizamos para aislar las variaciones estacionales:

$$\frac{Y_{i,t}}{T_{i,t} * e_{i,t}} = c_{i,t} * r_{i,t}.$$

En el caso aditivo lo aislamos con las diferencias

$$Y_{i,t} - T_{i,t} - e_{i,t} = c_{i,t} + r_{i,t}.$$

También podemos utilizar métodos específicos para eliminar el movimiento cíclico de una serie como el filtro de Hodrick y Prescot.

Una vez aislada la componente cíclica, esta se puede identificar mediante el análisis armónico y a partir de aquí estudiar su periodograma.

23.4 Análisis de la serie desde un punto de vista estocástico

Hasta ahora hemos analizado las componentes deterministas de la serie desde un punto de vista descriptivo o clásico. En esta sección vamos a analizar la parte residual o aleatoria desde un punto de vista estocástico.

La metodología de Box Jenkins requiere aislar la parte aleatoria de la serie quitando la tendencia y consiguiendo homocedasticidad mediante distintass transformaciones.

Box Jenkins desarrolló modelos estadísticos para series temporales donde la explicación de cada dato viene dada por su comportamiento anterior. Los modelos elaborados se denominan modelos ARIMA(p, d, q) (AutoRegresive Integrated Moving Average). Para poder aplicar estos modelos necesitamos al menos 50 observaciones.

Modelos ARIMA: son modelos paramétricos que utilizan fundamentalmente los coeficientes de autocorrelaciónn a la hora de analizar las propiedades de la serie temporal en términos de la interrelación temporal de sus observaciones.

Para identificar la estructura del modelo, Box Jenkins utiliza dos fuentes de información, la Función de Autocorrelación Simple (ACF) y la Función de Autocorrelación Parcial (ACF partial).

Coeficiente de autocorrelación (ρ_k): es la herramienta estadística que mide la correlación, es decir, el grado de asociación lineal que existe entre observaciones separadas k periodos. Estos coeficientes de autocorrelación proporcionan mucha información sobre como están relacionadas entre sí las distintas observaciones de una serie temporal.

En general ρ_k es la correlación de la variable Y_t y la variable Y_{t+k}, indicando cual es la relación de un valor de la serie sobre el que se produce k periodos después.

El coeficiente de autocorrelación de orden k viene dado por la expresión:

$$\rho_k = \frac{cov(Y_t, Y_{t+k})}{S^2(Y_t)}$$

Recordemos que los coeficientes de correlación no tiene unidades y están definidos entre -1 y 1.

Si $\rho_k = 0$, significa que los valores de la serie que distan k periodos son linealmente independientes.

Si $\rho_k = \pm 1$, entonces se puede afirmar que existe una dependencia lineal perfecta entre valores que distan k unidades.

Función de Autocorrelación Simple (ACF): es una función definida para los números naturales, donde a cada número natural k se le asigna el coeficiente de correlación ρ_k.

Así, la ACF se va a denotar como una sucesión de números $\rho_1, \rho_2, ...\rho_k$.

Como acabamos de ver, el primer término ρ_1 es la correlación de la variable Y_t y la variable Y_{t+1}. Esta nos indica la relación de cada valor de la serie con el siguiente.

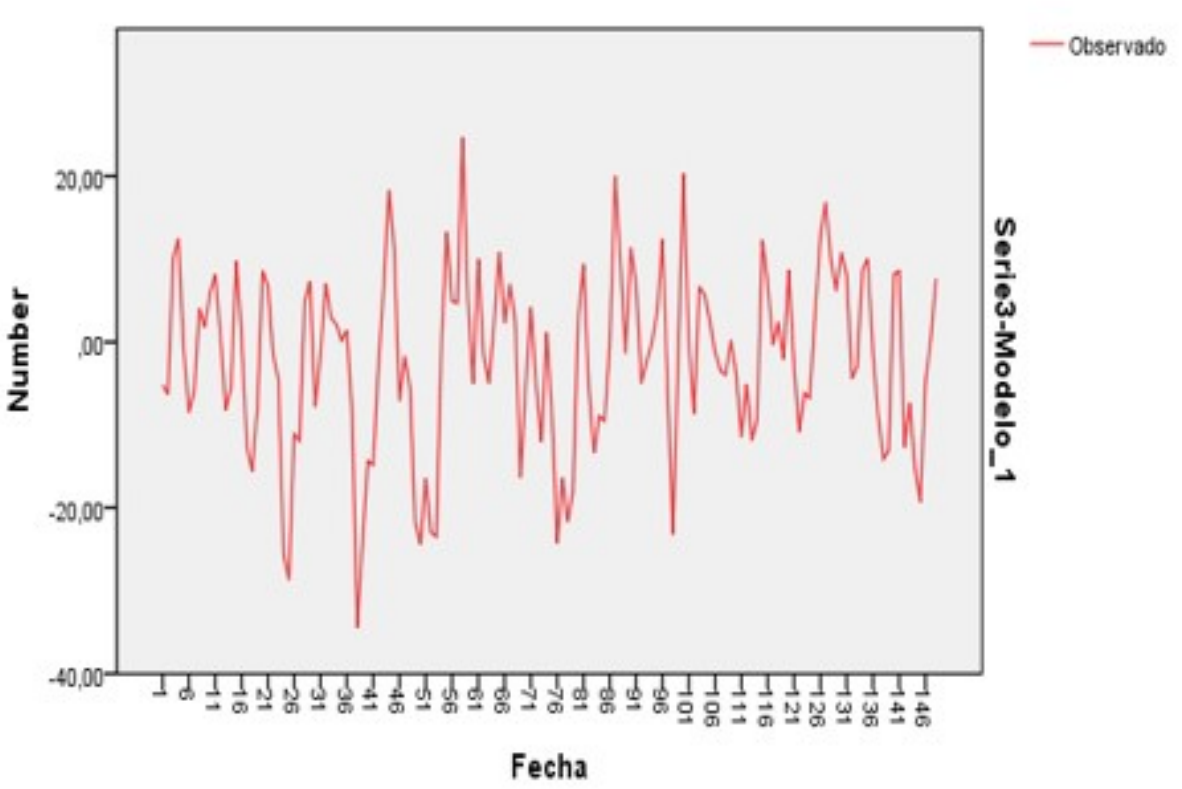


Figura 23.3: Representación gráfica de una serie temporal donde cada valor depende del anterior

Para construir la función de autocorrelacción utilizamos el sotfware IBM SPSS Statistics 27.

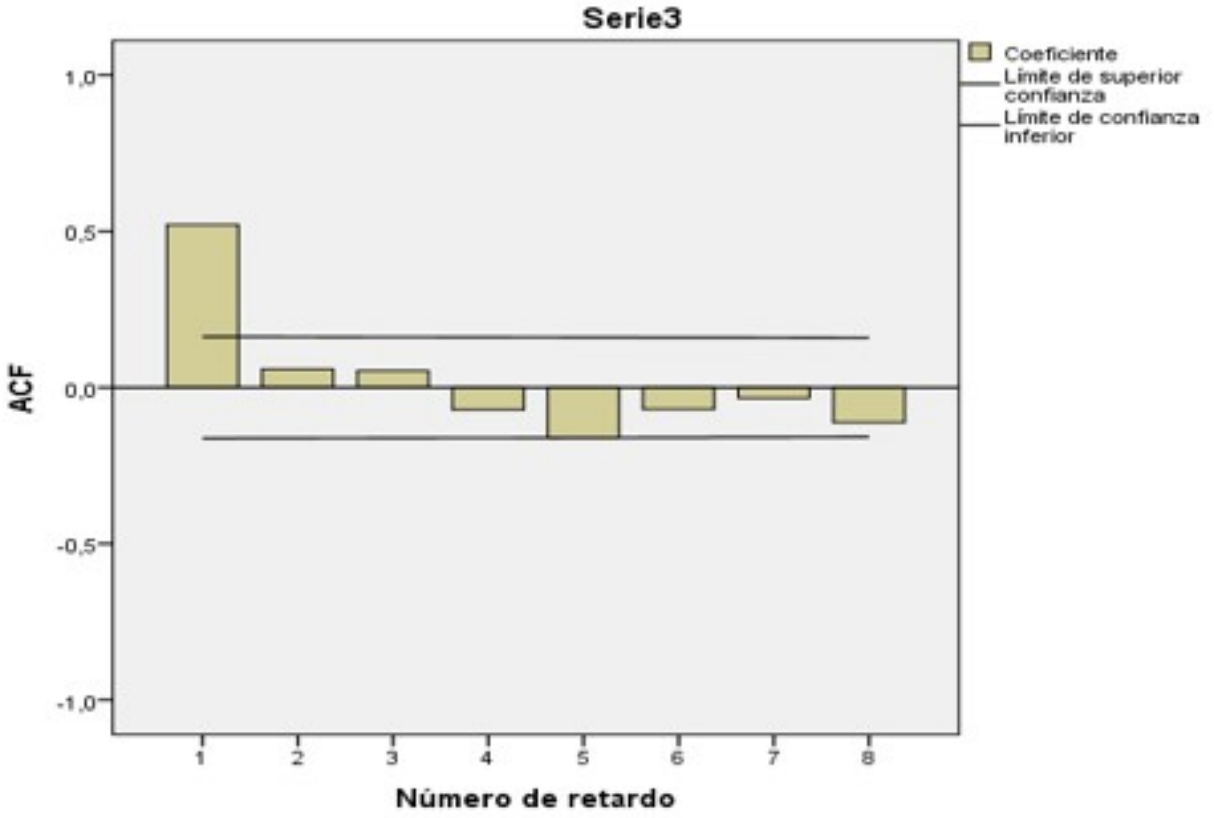


Figura 23.4: Representación gráfica de la función de autocorrelación

En al Figura 23.4 se puede ver como los datos de esta serie se pueden calcular a partir del valor en el periodo anterior, el resto de los valores aportan muy poca información ya que el resto de los coeficientes de autocorrelación no son significativos.

Si se analiza con detalle el significado de la función de autocorrelación concluiremos que Y_{t+2} depende de Y_{t+1} y este a su vez depende de Y_t, por lo tanto se puede concluir que Y_{t+2} depende de Y_t.

Para saber si la dependencia que existe entre dos observaciones que distan k periodos es por la influencia de valores intermedios o es directamente por el valor que dista k periodos necesitamos conocer la función de autocorrelación parcial.

Función de Autocorrelación Parcial: La función de autocorrelación parcial proporciona la relación directa que existe entre observaciones separadas por k retardos. Esta es una información muy valiosa sobre la estructura de la serie, ya que elimina el problema que presentaba la función de autocorrelación simple.

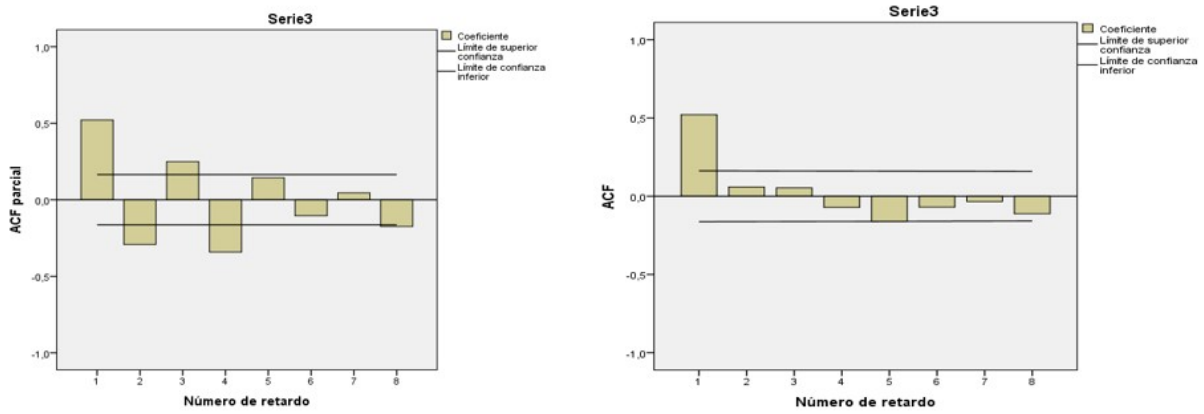


Figura 23.5: Representación de la función de autocorrelación y de la función de autocorrelación parcial.

Si se comparan ahora las funciones de autocorrelación y la de autocorrelación parcial de la Figura 23.5 se puede determinar el modelo de Box Jenkins que sigue la serie temporal. En la función de autocorrelación solamente es significativo el coeficiente de orden 1, es decir cada valor de la serie solo depende del anterior, sin embargo, por la propiedad transitiva, también puede depender de otros valores del pasados. Para comprobar si existen más valores del pasado que influyen en la serie temporal nos fijamos en las autocorrelaciones parciales y aquí se observa que el presente también depende de lo que ocurra en 2, 3 y cuatro periodos anteriores.

Las series temporales se pueden clasificar en dos grupos. Las primeras son las que responden a procesos estables en el tiempo, por ejemplo la cantidad de lluvia en un determinado municipio a lo largo de los años, la temperatura media, etc. Estos procesos se denominan estacionarios. El otro grupo de series son aquellas que no son estacionarias, bien porque presenten una tendencia no constante o variaciones estacionales. Entre este grupo de series está el Producto Interior Bruto de un país, el precio del oro, etc. Para poder clasificar una serie temporal en uno de los grupos debemos fijarnos en el periodo de estudio. Una determinada serie puede ser estable a corto plazo pero no serlo en un periodo bastante más largo.
Los modelos para identificar una serie temporal son diferentes si la serie es estacionaria o no, por lo tanto lo primero que debemos hacer es comprobar si se puede conseguir transformar la serie temporal en un proceso estacionario mediante algún tipo de transformación.

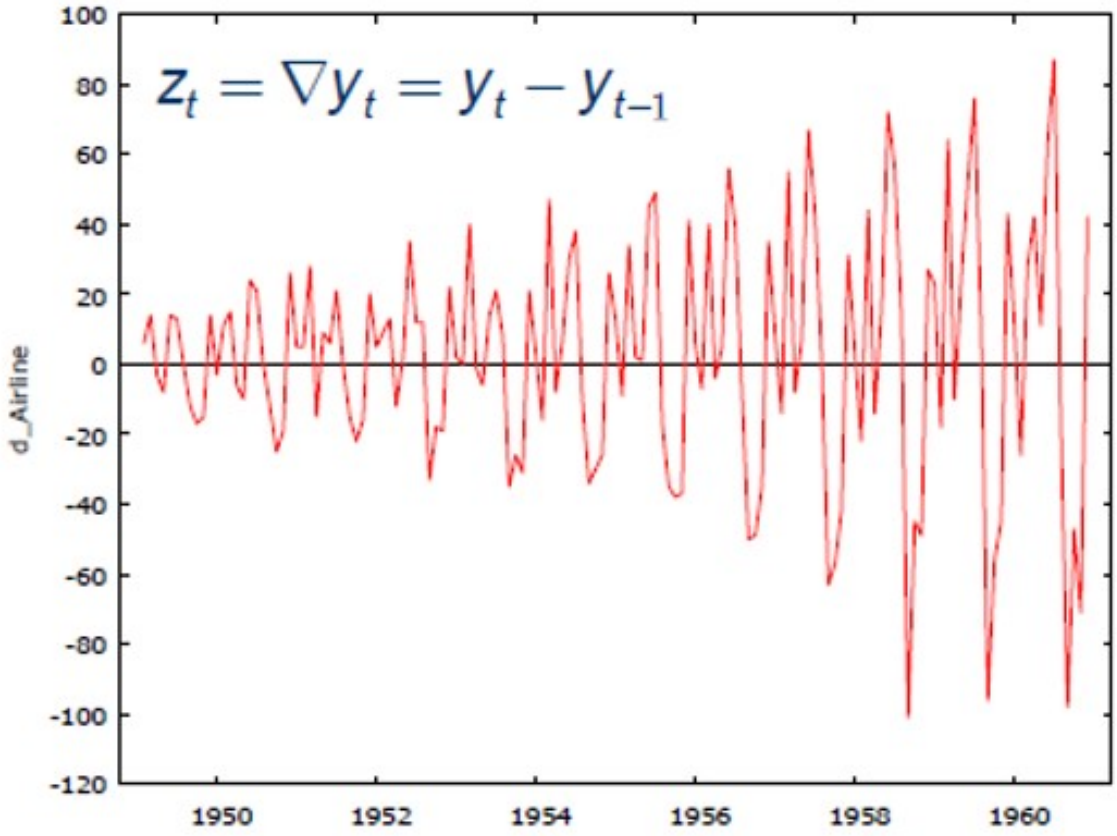


Figura 23.6: Representación de una serie temporal con heterocedasticidad.

La transformación más utilizada para conseguir estacionar una serie temporal es la diferenciación de la serie $(Z_t = Y_t - Y_{t-k})$, aunque en ocasiones con esta transformación no se consigue estabilizar la varianza. Si se observa la Figura 23.6 se comprueba que al diferenciar la serie temporal se ha conseguido que la media se mantenga constante en el tiempo, sin embargo la varianza va aumentando con el tiempo (heterocedasticidad). Ahora se debería de realizar otra transformación para conseguir homocedasticidad.

Para conseguir una serie estacionaria en algunas ocasiones se debe combinar la transformación logarítmica con la diferenciación $Z_t = \ln(Y_t) - \ln(Y_{t-k})$.

23.4.1 Procesos autoregresivos

En el análisis de series temporales el objetivo es utilizar la teoría de procesos estocásticos para determinar que caracteriza el comportamiento de la serie y cómo predecir en el futuro. Si se quieren conseguir métodos de predicción consistentes, no se puede utilizar cualquier tipo de proceso estocástico, sino que es necesario que la estructura probabilística del mismo sea estable en el tiempo.
Puesto que se va a predecir el futuro a partir del pasado, necesitamos que la serie tenga algún tipo de estabilidad.

Los modelos más simples para los procesos estacionarios son los modelos autorregresivos, que generalizan la idea de regresión para representar la dependencia lineal entre dos variables aleatorias.

$$Y_t = \mu + \varphi X_t + r_t$$

La función de autocorrelación de un proceso estacionario es una función de k que normalmente se representa mediante un gráfico de barras que se denomina **correlograma**.

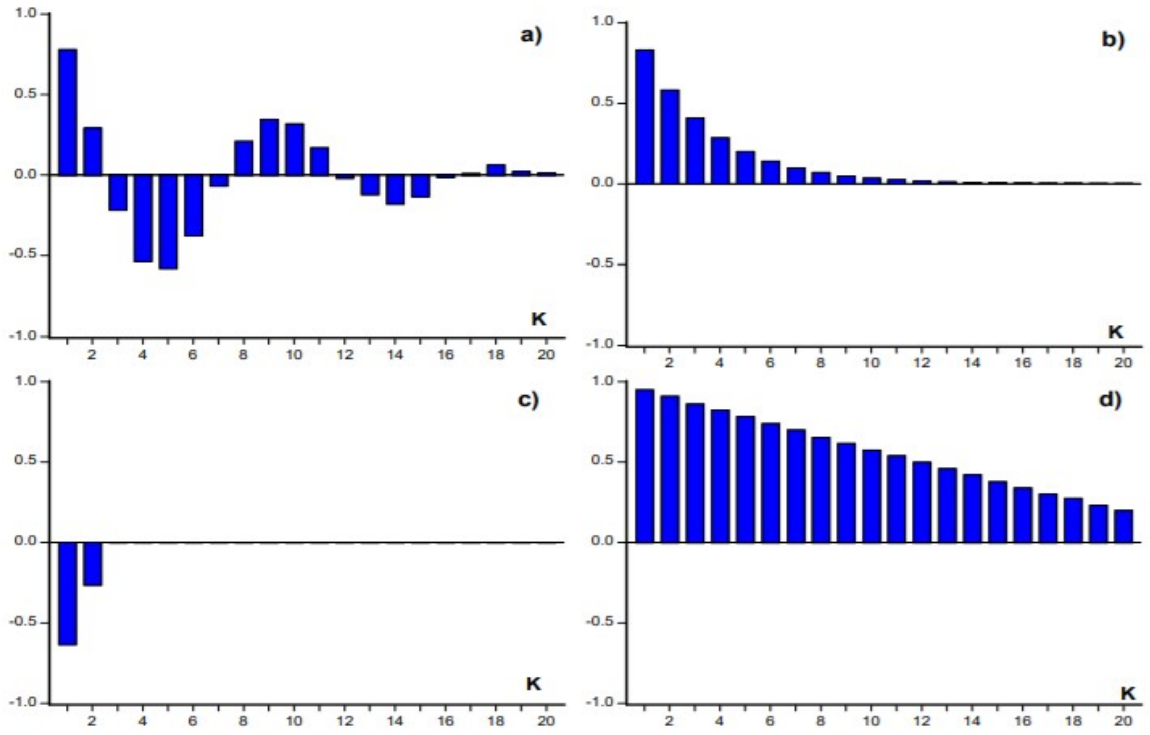


Figura 23.7: Representación de distintos correlogramas de procesos estacionarios (Fuente: González Casimiro 2009).

Los gráficos a), b) y c) de la Figura 23.7 corresponden a distintas series estacionarias puesto que las autocorrelaciones decrecen de forma exponencial a medida que aumenta k mientras que el gráfico d) pertenece a una serie no estacionaria, esto se observa en el lento decrecimiento de las autocorrelaciones . En el caso a) hay alternancia de autocorrelaciones positivas y negativas

Modelo de ruido blanco

El modelo más sencillo sería el ARIMA(0,0,0), este modelo no tiene parte regular ni parte estacional, habitualmente se le conoce como ruido blanco (r_t). Es una secuencia de variables aleatorias de media cero, varianza constante y autocorrelaciones nulas.

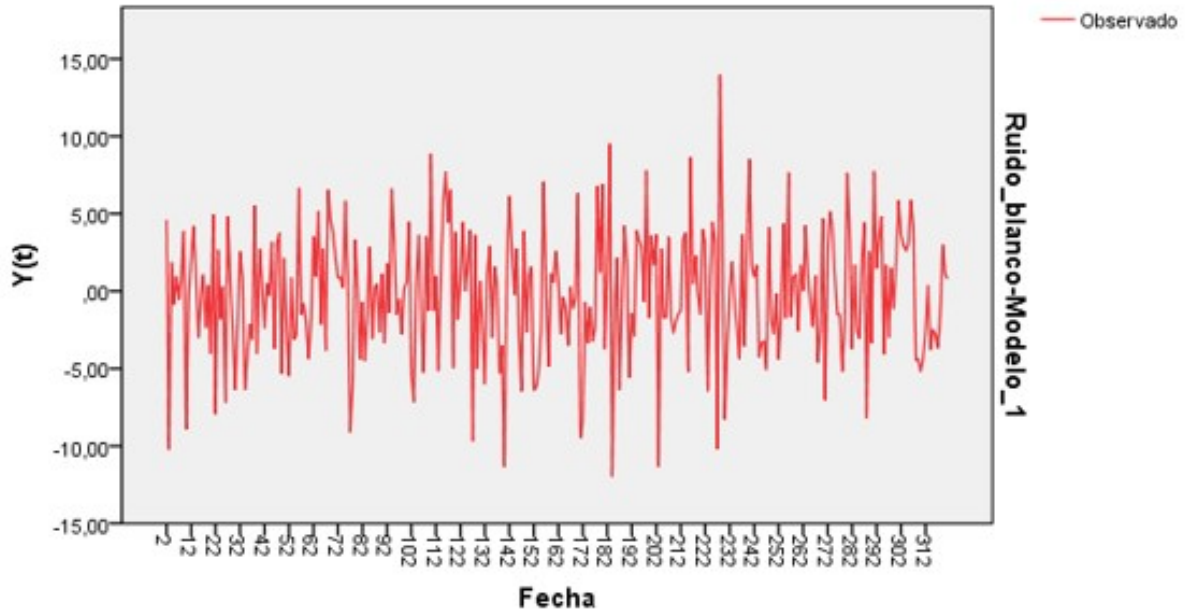


Figura 23.8: Representación de un ruido blanco.

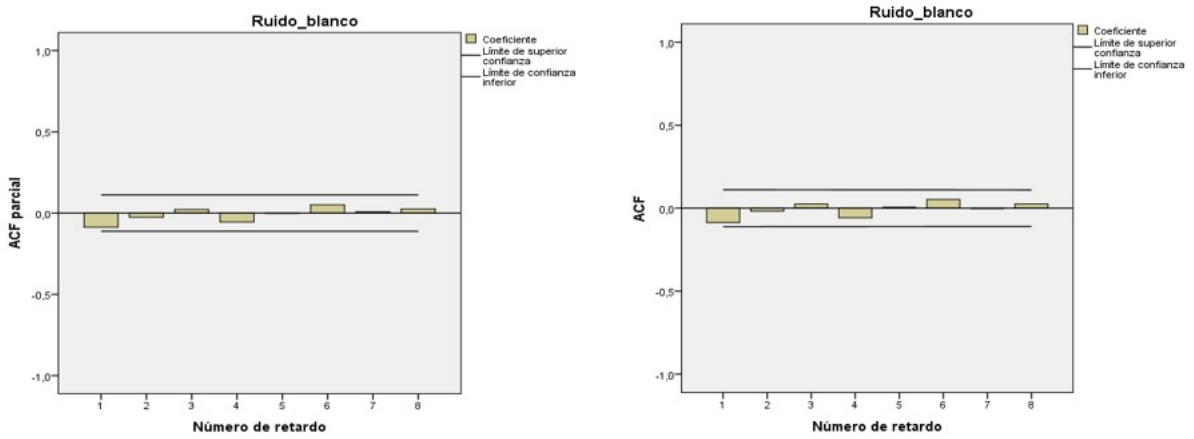


Figura 23.9: Representación de las funciones de autocorrelación parcial y autocorrelación de un ruido blanco.

En la Figura 23.8 se ha representado una simulación de un ruido blanco, se observa que la serie oscila entorno a la media (0) y con varianza constante, sin presentar ningún patrón, es decir la autocorrelación entre valores que distan k periodos es prácticamente nula como se observa en la Figura 23.9. Dado que uno de los objetivos es predecir los valores del futuro, en el caso de un ruido blanco, el futuro no depende del pasado.

Modelos AR(k)

Modelo AR(1) El modelo ARIMA(1, 0, 0) se el denota también como AR(1) es aquel modelo que solo depende de la observación anterior y una perturbación aleatoria (ruido

blanco). Un ejemplo puede ser la cantidad de agua de un embalse.

$$Y_t = \mu + \varphi Y_{t-1} + r_t$$

donde $|\varphi| < 1$

Se trata de utilizar la información de la función de autocorrelación (ACF) y la autocorrelación parcial (ACF partial) para definir un patrón que reproduzca el comportamiento de la serie.

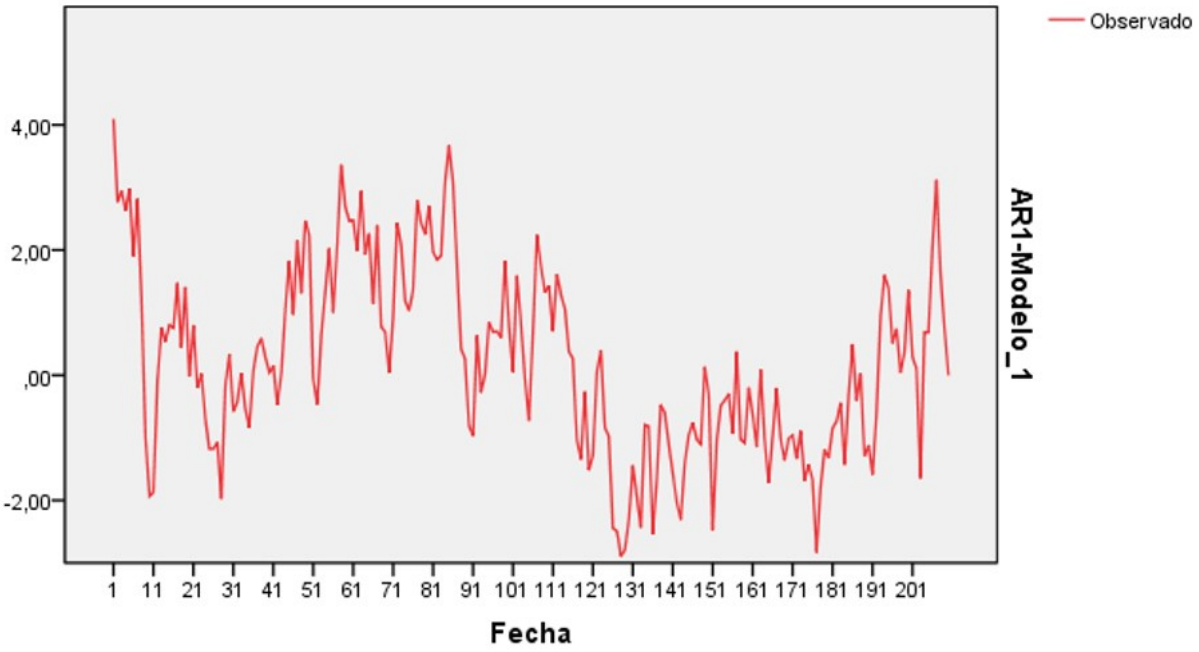


Figura 23.10: Representación de un modelo AR(1).

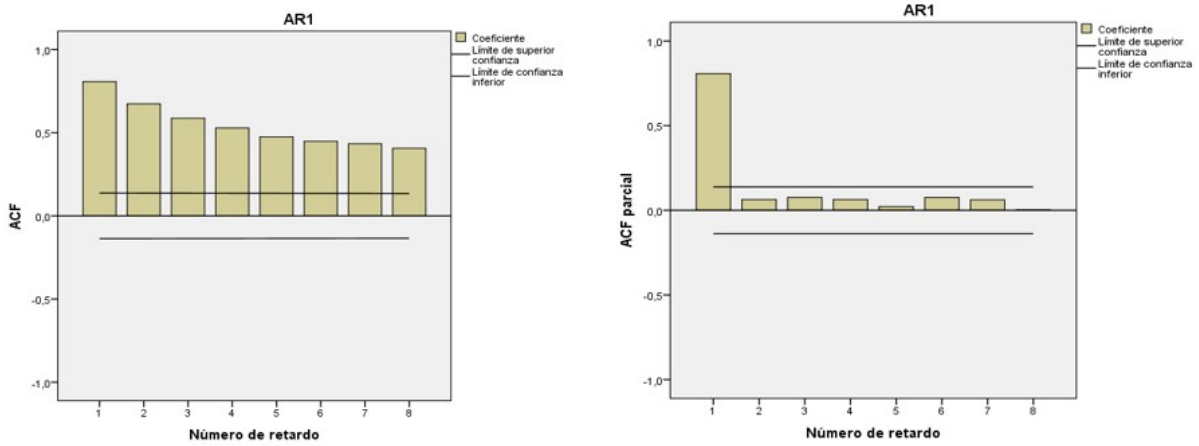


Figura 23.11: Correlogramas de un modelo AR(1).

Como se observa en la Figura 23.10, en la función de las autocorrelaciones hay más de un coeficiente significativo, lo que significa que Y_t no solo depende de la observación

anterior, también depende de la que dista 2, 3 o 4 periodos. Sin embargo, si observamos las correlaciones parciales, tan solo el de orden 1 es significativo, por lo tanto la influencia directa de las observaciones que distan 2, 3, o 4 periodos es cero.
El correlograma del modelo AR(1) con parámetro φ positivo indica que tienen todos los coeficientes de autocorrelación positivos con decrecimiento exponencial. Es una serie cuyo comportamiento en media es estable entorno al 0 su dispersión es estable y con rachas de observaciones por encima de la media seguidas de otras por debajo de la media. Si el parámetro del modelo φ fuese negativo, los coeficientes de autocorrelación alternarían el signo y su comportamiento sería mucho más ruidoso.

Se puede demostrar que cuando tenemos un modelo AR(1) los coeficientes de autocorrelación son de la forma: $\rho_k = \varphi^k$

Modelos AR(2)

Los modelos AR(2) tienen la siguiente forma:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + r_t$$

siendo r_t un ruido blanco. Además, las raíces del polinomio autorregresivo

$$0 = 1 - \varphi_1 x - \varphi_2 x^2$$

deben estar fuera del círculo unidad.

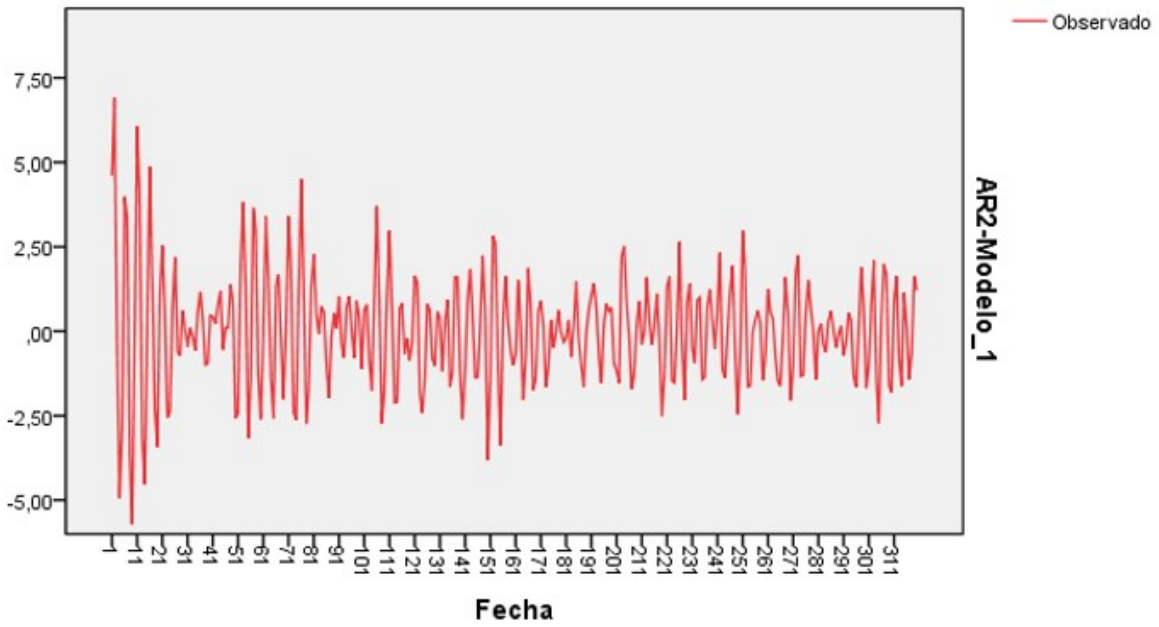


Figura 23.12: Representación de un modelo AR(2).

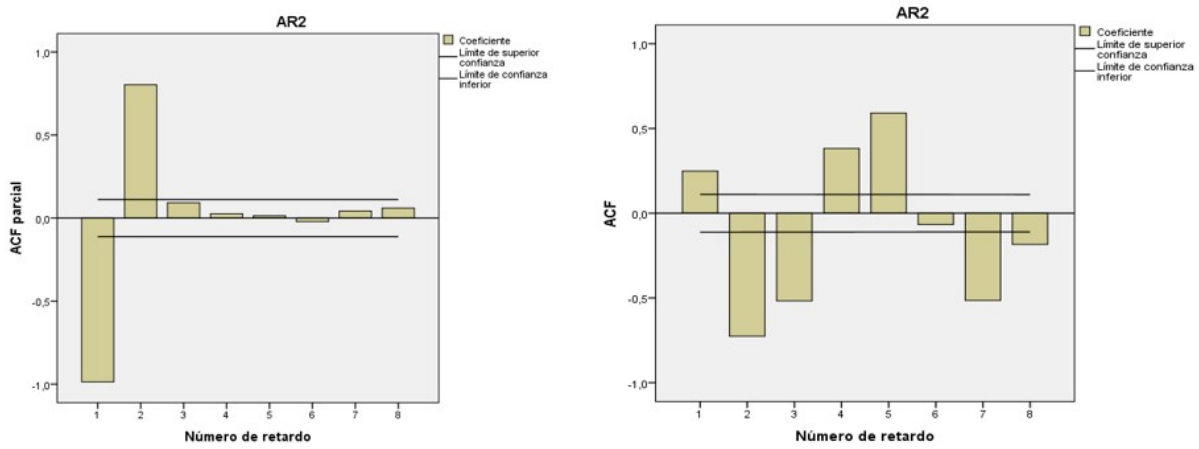


Figura 23.13: Representación de las funciones de autocorrelación parcial y autocorrelación de un modelo AR(2).

Los modelos AR(2) no presentan una única estructura en su función de autocorrelación, cuando las raices del polinomio autoregresivo son reales, la función de autocorrelación decrece exponencialmente con todos los los coeficientes positivos o con alternancia en el signo. Si las raices del polinomio autoregresivo son complejas entones la función de autocorrelación decrece exponencialmente pero con forma de onda seno-coseno como es el caso de la Figura 23.13 donde el correlograma de la función ACF tiene forma senoidal,

Modelo AR(k)

En general, los modelos AR(k) se pueden expresar como:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \varphi_k Y_{t-k} + r_t$$

siendo r_t un ruido blanco y el polinomio autoregresivo tendrá k raices raices, en general, distintas.

Determinar el orden de un modelo autorregresivo a partir de la función de autocorrelación es difícil, ya hemos visto que estas funciones son una mezcla de decrecimientos exponenciales y sinosoidales. Este es el motivo por el que para los modelos AR(1) y AR(2) se representaba el correlograma de la función de autocorrelación parcial.
Si comparamos el modelo AR(1) con AR(2) las funciones de autocoorelación pueden ser muy parecidas pero las funciones de autocorrelación parcial son diferentes, mientras que el el modelo AR(1) el efecto de Y_{t-2} sobre Y_t es siempre a través de Y_{t-1}, en el modelo AR(2) además del efecto de Y_{t-2} que se transmite a Y_t a través de Y_{t-1}, existe el efecto directo de Y_{t-2} sobre Y_t.
Luego para que una serie temporal se ajuste a través de un modelo AR(k) no solo la función de autocorrelación debe ser una mezcla de decrecimientos exponenciales y sinosoidales a medida que aumenta la distancia entre las observaciones, además la función de autocorrelación parcial tendrá los k primeros coeficientes significativamente distintos de cero.

23.4.2 Procesos de medias móviles

En la sección anterior se han analizado algunos modelos estacionarios, los procesos autoregresivos. Estos modelos tienen memoria relativamente larga, ya que el valor actual depende de todos los anteriores, aunque los coeficientes de autocorrelación vayan decreciendo de forma exponencial.
Si nuestra serie presenta memoria muy corta, es decir una innovación en la serie genera un número muy pequeño de perturbaciones no puede modelizarse mediante un modelo AR(k). Sin embargo los modelos de medias móviles MA(k) tienen memoria mucho más corta, una acción exterior solo produce perturbaciones en un pequeño número de periodos.
El modelo MA(k) es una aproximación al modelo lineal general.

$$Y_t = \mu + r_t - \varphi_1 r_{t-1} - \varphi_2 r_{t-2} - ...\varphi_k r_{t-k}$$

siendo r_t es un ruido blanco.

Modelos MA(1)

El modelo ARIMA(0,0,1) se le denota como MA(1), media móvil del orden 1.

$$Y_t = \mu + r_t - \varphi r_{t-1}$$

donde $|\varphi| < 1$ *y* r_t es un ruido blanco.

Al analizar la estructura de un procesos de medias móviles MA(1) la perturbación r_t aparece en el sistema en el momento t e influye en Y_t y en Y_{t+1} únicamente, por lo que su memoria es de un solo periodo

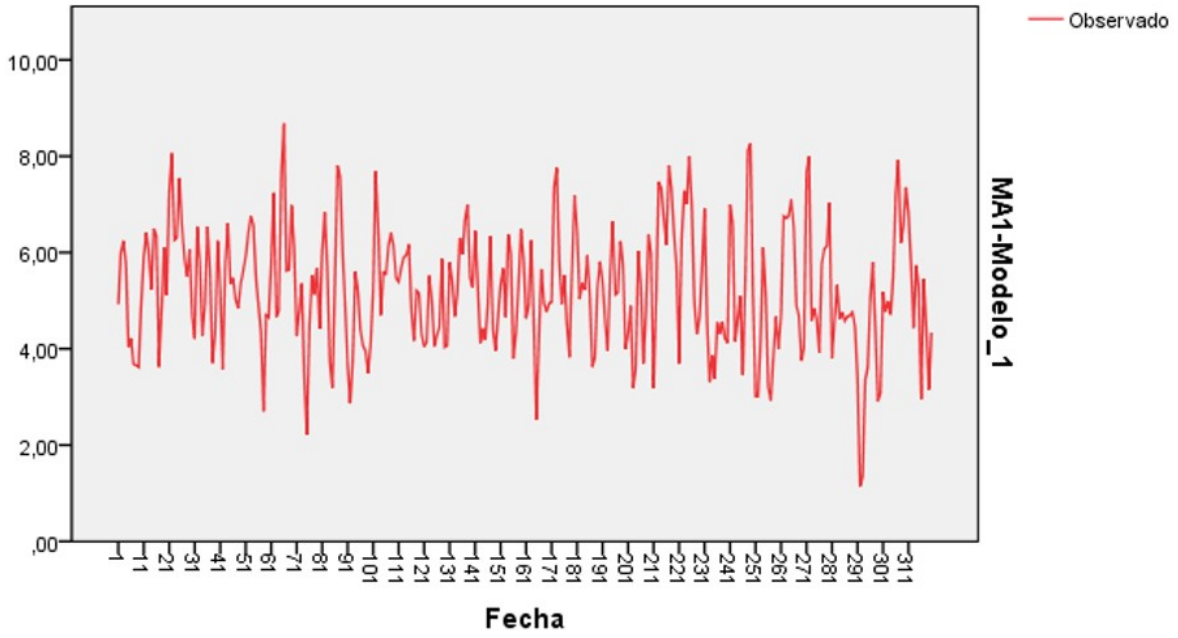


Figura 23.14: Representación de un modelo MA(1).

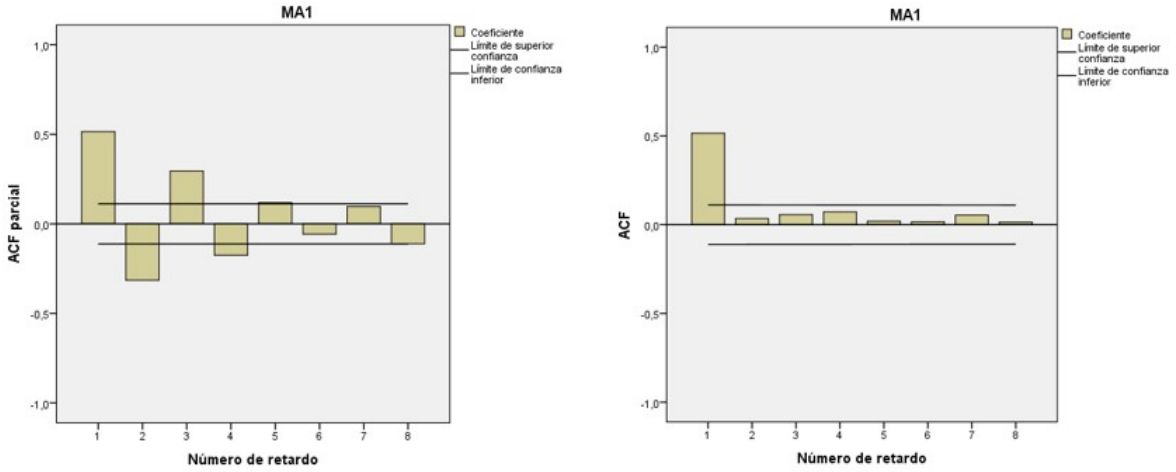


Figura 23.15: Representación de las funciones de autocorrelación parcial y autocorrelación de un modelo MA(1).

Para los modelos MA(1) solamente el primer coeficiente de la función de autocorrelación es distinto de cero, pudiendo ser ρ_1 positivo o negativo. En la Figura 23.15 se observa que $\rho_1 > 0$ lo que nos indica que la serie oscila en torno a la media con una varianza constante y bastante suave. En el caso en el que $\rho_1 < 0$ la serie temporal sería menos suave.

Si comparamos el modelo MA(1) con el AR(1) se observa que el gráfico de la función de autocorrelación (ACF) del modelo MA(1) coincide con el gráfico de la función de autocorrelación parcial (ACF parcial) del modelo AR(1), y viceversa. El modelo MA(1) tiene mucha menor memoria que el modelo AR(1).

En general el modelo MA(k) tiene unas características similares al MA(1), sin embargo, cualquier acción exterior influye en los k periodos posteriores de la serie temporal.Es decir tiene mayor memoria y la estructura representada por el modelo es más rica. Cualquier perturbación r_t influye en Y_t, Y_{t+1} y en el resto de los valores hasta Y_{t+k}, permaneciendo dicha alteración en el modelo k periodos.

23.4.3 Procesos integrados

Normalmente la series no son procesos estacionarios en media y en varianzas. Diremos que una serie sigue un proceso integrado de orden q si Y_t no es estacionario, pero su diferencia de orden q sigue un proceso estacionario e invertible.
Una propiedad importante que diferencia los procesos integrados de los estacionarios es la forma en la que desaparece la dependencia con el tiempo. En los procesos estacionarios las autocorrelaciones disminuyen exponencialmente con el tiempo, haciéndose practicamente cero en un número pequeño de retardos. En los procesos integrados las autocorrelaciones disminuyen linealmente en el tiempo y es posible encontrar coeficientes de autocorrelación significativamente distintos de cero para reatardos muy altos.

Paseos aleatorios

El modelo ARIMA(0,1,0) es un proceso estocástico no estacionario que se denomina paseo aleatorio y viene dado por la expresión

$$Y_t = \mu + Y_{t-1} + r_t$$

Observemos que un modelo ARIMA(0,1,0) es equivalente a un modelo AR(1) con $\varphi = 1$. Como hemos visto anteriormente el modelo AR(1) viene dado por la expresión

$$Y_t = \mu + \varphi Y_{t-1} + r_t$$

donde $|\varphi| < 1$

Si $|\varphi| > 1$, la serie sería explosiva y en el infinito valdría infinito y por lo tanto no es una serie interesante de analizar. Si $|\varphi| = 1$, la serie no es estacionaria pero tampoco es explosiva y pertenece a la clase de procesos integrados de orden uno.

$$w_t = \bigtriangledown Y_t = \mu + r_t$$

es un proceso estacionario. A este proceso se le llama paseo aleatorio.

Especial interés tiene μ así como en los modelos estacionarios la inclusión de una constante solo afecta al nivel promedio de la serie, en los modelos no estacionarios sus efectos son más importantes. La inclusión de una constante en un modelo no estacionario implica que la serie tienen una tendencia con pendiente μ además de la tendencia estocástica

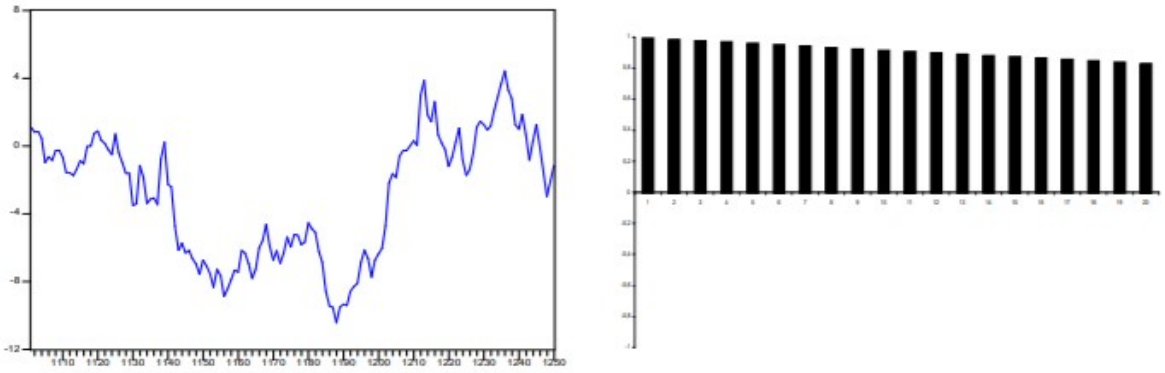


Figura 23.16: Paseo aleatorio con $\mu = 0$ (Fuente: González Casimiro 2009).

En la Figura 23.16 se ha simulado un paseo aleatorio. Se puede observar que la evolución de la serie no tiene las características de los procesos estacionarios. La serie se mueve aleatoriamente por encima y por debajo de la media. Lo que nos indica que puede ser un paseo aleatorio es el correlograma, los coeficientes de aurtocorrelación decrecen muy lentamente.

Podemos encontrar más modelos en: Mauricio 2007

Bibliografía

Peña, Daniel (2005). *Análisis de series temporales*. Alianza (página 153).
Mauricio, José Alberto (2007). "Análisis de series temporales". En: *Universidad Complutence de Madrid* (página 172).
González Casimiro, María Pilar (2009). "Análisis de series temporales: Modelos ARIMA". En: (páginas 153, 165, 172).
Villagarcía, Teresa (2018). "Series temporales". En: *Obtenido de http://www. est. uc3m. es/esp/nueva_docencia/leganes/ing_industrial/estadistica_industrial/d oc_grupo1/archivos/Apuntes % 20de % 20series. pdf* (página 153).